THE SMALL-WORLD NETWORK OF SQUASH

BY MICHAEL KEARNS AND RYAN RAYFIELD

Not all social networks are built in front of glowing monitors with a Mountain Dew and a bag of Cheetos at hand. There are some social networks in which participation is outright good for your health—like squash. Using tools from the emerging field of network science we will investigate the specialized social network in which each node is a squash player, and there is a link between any pair of players who have played a match before.

The source data for our study was all US Squash singles matches recorded over a recent multi-year period. The number of players in this network was 26,503 and the number of matches was 240,446. The average number of matches played per player was 18.4 and the maximum was 210 (by Gabriel Bassil of Brooklyn). Like virtually all large-scale social networks, the squash network is sparse, meaning that the number of matches actually played was only a tiny fraction of those possible—less than 7 hundredths of 1 percent. It was also the case that a small number of the most active players account for a disproportionate fraction of the total matches; in network science parlance, the distribution of the number of matches across players is heavy-tailed.

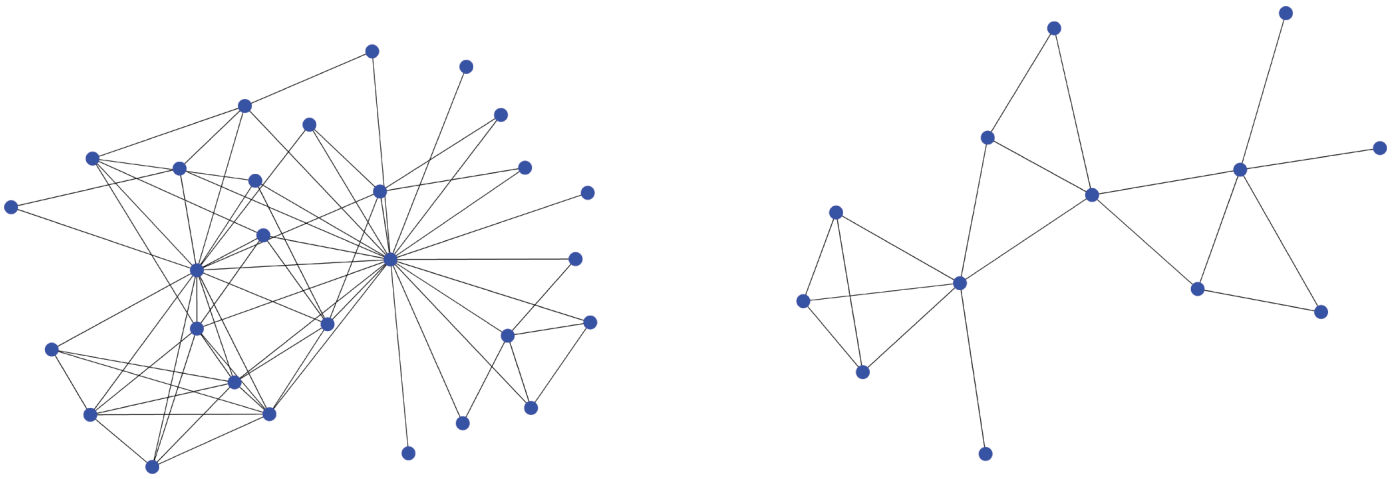
To understand the global shape or structure of our network, we need to examine the *connected components*, which are the islands of connectivity. Let's consider two players as living in the same island if there is any chain of matches that connects them. So if Alice played Bob, and Bob played Charlie, and Charlie played Dana, then Alice and Dana are in the same connected component (or "island") by virtue of this chain, even if they have never played each other.

Network science predicts that in any real social network, there should be a giant component—a mainland which contains the vast majority of the population—along with an archipelago of much smaller islands with no links to the mainland. This was the case with our data. The largest component of the squash network contained almost 99% of the players. Intuitively it's hard for two large components to coexist: all it takes is one match between a player from each island and the two merge to become one.

What about the 1% of players in the archipelago, which consists of 77 additional components? What do these tiny islands look like? Unlike Facebook, playing squash requires physical proximity, so it is not surprising that many of the tiny components had a strongly geographic flavor. For instance, the second largest component had only twenty-eight players, all of whom live in Raleigh, NC, while the third largest consisted exclusively of players in San Antonio. Many of the other small components were lonely, isolated pairs of players who had only played each other. We encourage them to play more squash and join the giant component.

Not all the players in the giant component are

Visualization of the "mainland" of the US Squash network.



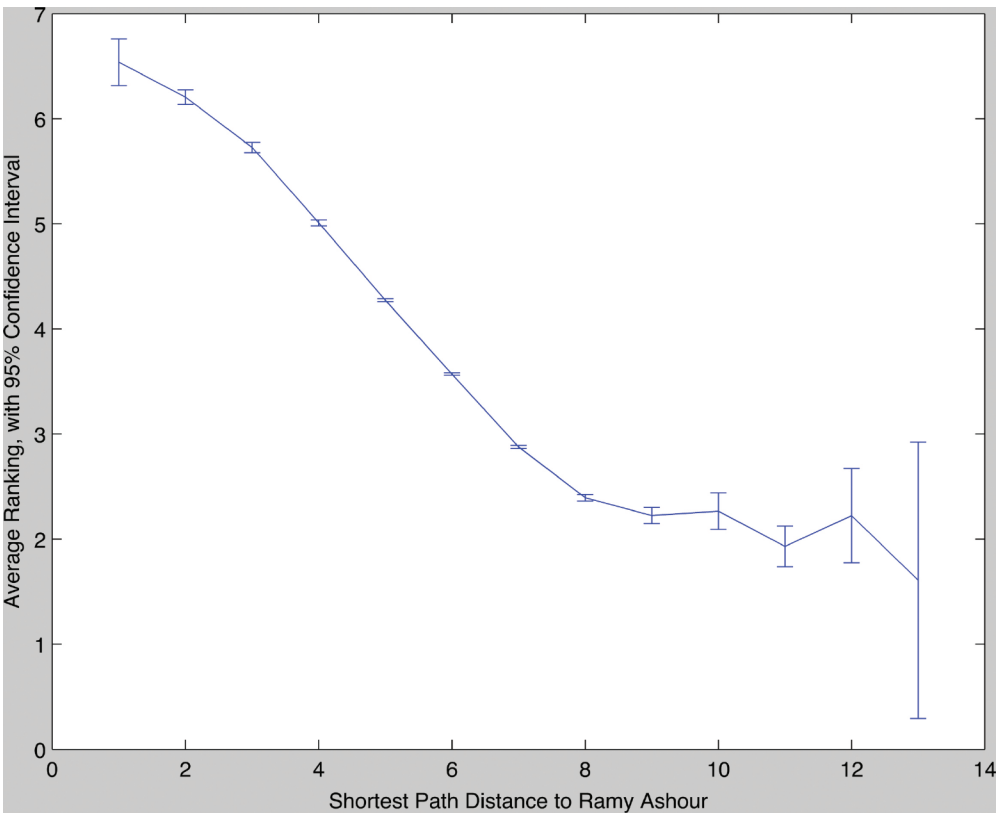
Visualization of the Raleigh and San Antonio connected components or "islands".

necessarily connected by short chains of matches. Indeed, the longest shortest chain between two players in the giant component was the nineteen-match chain that connected Simon Anderson of Madison, Wisconsin to Ari Wolgin of Philadelphia. But overall, the small world property does indeed hold for the squash network: on average, a chain of only 4.8 matches (or degrees of separation) connected a typical pair of players in the giant component.

Speaking of distances, it can be illuminating to consider shortest chains to particular players of interest. Some readers may be familiar with the popular parlor game "Six Degrees of Kevin Bacon," but since Kevin is not (yet) a squash player, we'll instead use Ramy Ashour. Let's define your Ashour number to be the length of the shortest chain of matches connecting you to the great Egyptian world champion. So if you're Ramy Ashour, your Ashour number is zero. If you have played a match with Ramy Ashour, your Ashour number is one. If you haven't played a match with him, but have played someone who has, your Ashour number is two. And so on.

For example, coauthor Kearns has a twelve-year old son named Gray with an Ashour number of seven, via the following chain of matches: Gray Kearns—Ben Stewart—Auggie Bhavsar—Jimmy Li—Ryan Rayfield—Chris Hanson—Alan Clyne—Ramy Ashour. Note that in another demonstration of small worlds, coauthor Rayfield appears along this path.

It turns out that player ratings increase steadily with each hop along this chain from Gray to Ramy. This is far from a fluke.On average, the higher your Ashour number, the lower your rating. (In precise statistical terms, the correlation between Ashour numbers and ratings is strongly negative, -0.76; a correlation of -1 would mean your Ashour number completely determines your rating.) In other words, Ashour numbers are actually already a pretty good rating system *even though they entirely ignore the outcome of any matches* and only measure a kind of social distance. This is a consequence of the broad fact that our network strongly exhibits what sociologists call homophily: the concept that birds of a feather flock together. In our case this means that there is a strong bias in the network towards similarly skilled players playing each other.

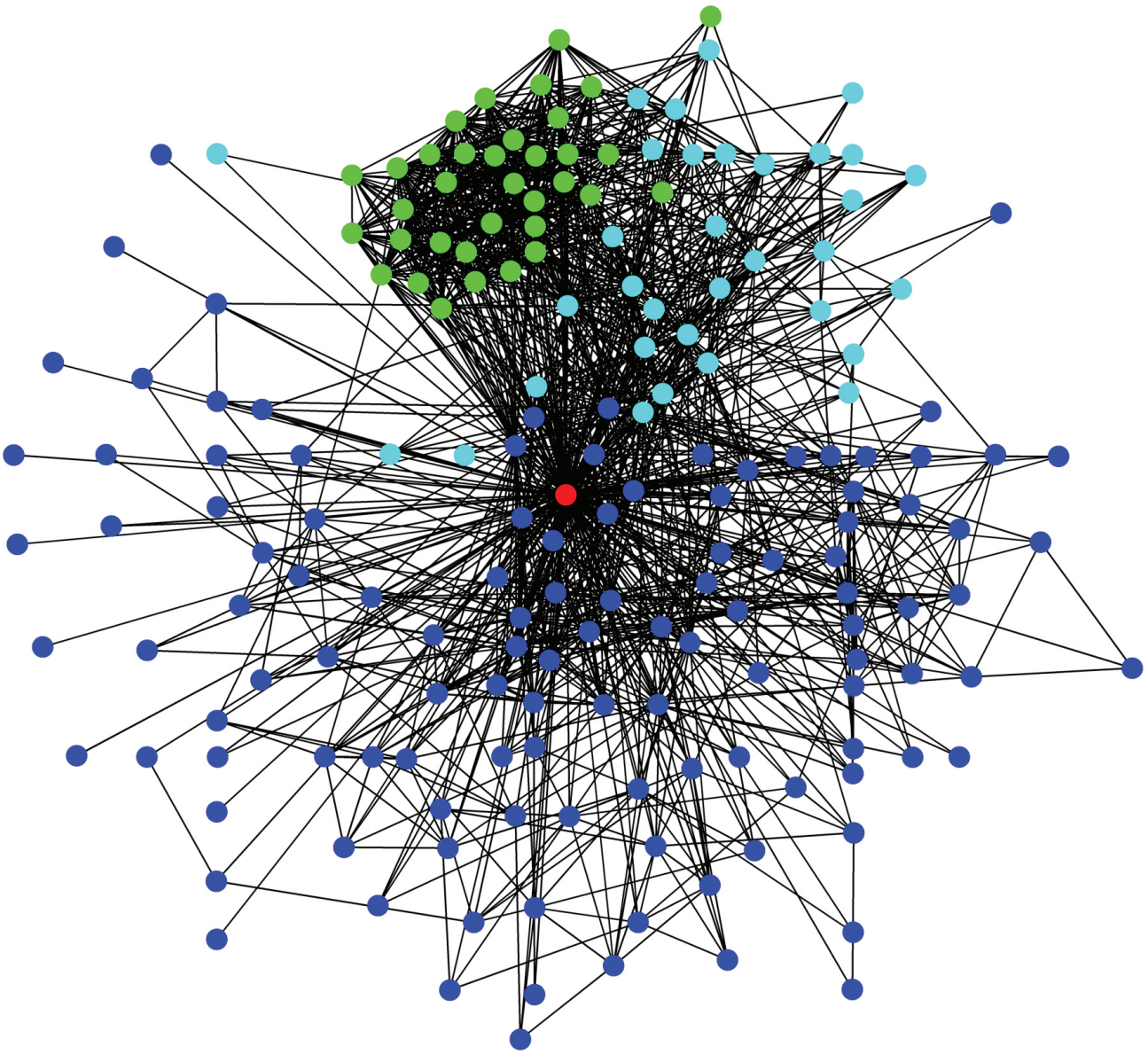


Distance to Ramy Ahsour vs. average player rating.

While Ramy Ashour is certainly an important player, in ne twork terms he lives in an elite and remote neighborhood that few of us will ever visit. But there are other notions of importance in network science that capture being in the middle of the network rather than in an enclave. One of these notions is known as *betweenness centrality*. This measures how many chains between *other* pairs of players pass through you, and thus illuminates the extent to which you are in the middle of the network, or a hub of traffic.

Unlike the players with very small Ashour numbers, players with higher centrality tend to have moderately strong rather than stratospheric ratings. These players tend to play a lot but, more importantly, they seem to play a diverse collection of opponents, both in ratings and geography.

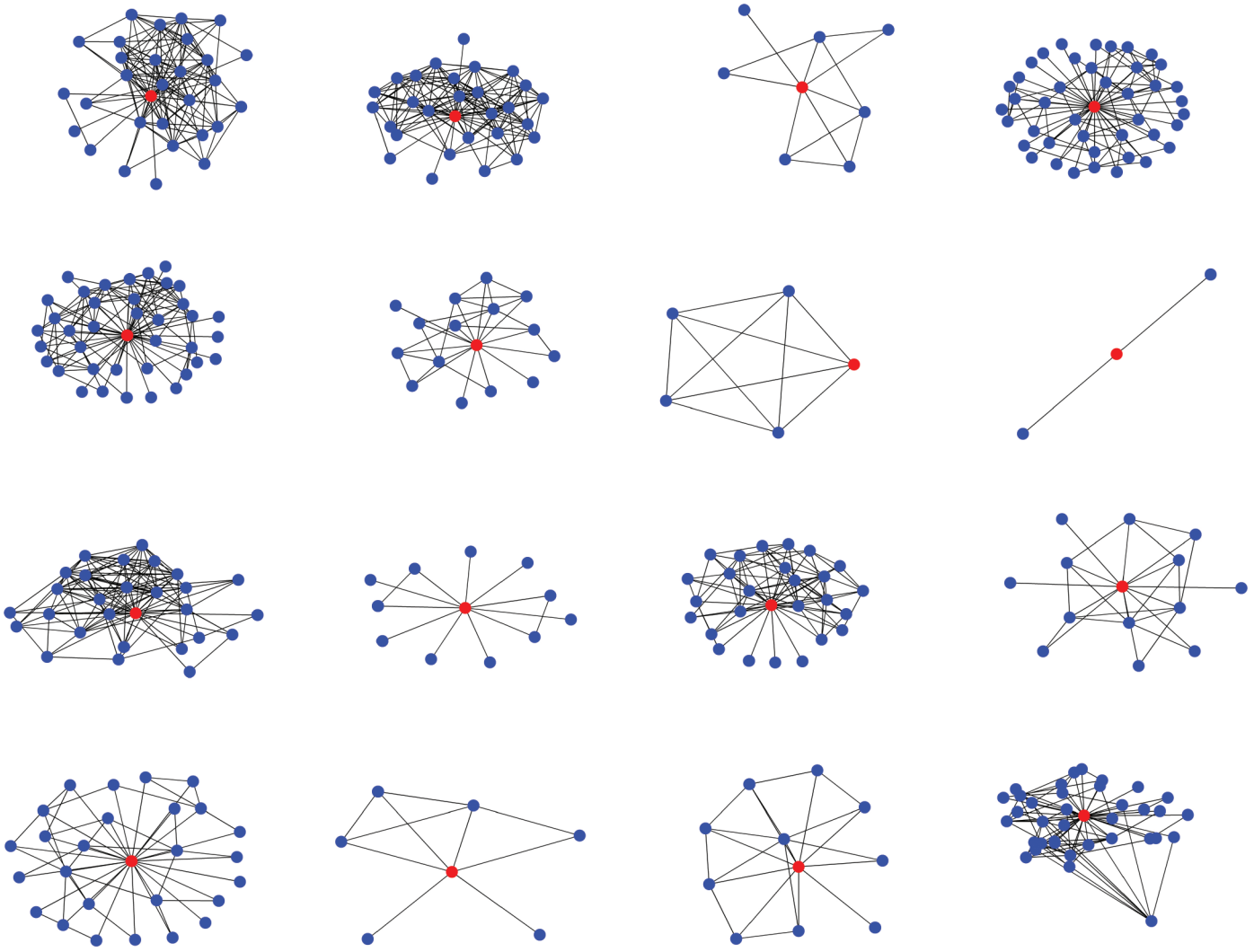
The MCP (Most Central Player) Award went to Dillon Huang, a junior player from Fremont, CA. About 2.5% of all shortest paths in the entire network passed through Dillon—roughly 650 times what you'd expect if the network consisted of entirely randomly chosen matches. In the accompanying visualization of his local or "ego" network, Dillon is shown as a red node in the center of his past opponents. We ran a standard clustering algorithm on Dillon's network, dividing his opponents into color-coded groups that played amongst themselves a great deal. The algorithm found three relatively distinct clusters, reflecting the fact that Dillon was a top junior who tended to enter tournaments for higher age brackets, as well as adult open tournaments. Who knows—as the squash network evolves, perhaps in a few years we will be discussing Huang numbers rather than Ashour numbers.



Local network for most central player Dillon Huang. Colors indicate groups of Dillon's opponents who have played against each other frequently.

What do the squash neighborhoods of mere mortals look like? Below we show the local networks for a handful of randomly selected players. In each case the sampled player is shown as a red node, all of their past opponents as blue nodes, and all matches between opponents are shown as links.

Perhaps our study and the Ashour number will inspire you to alter your network structure by playing more matches and playing with opponents you might not have otherwise—maybe even with Ramy Ashour himself.



The variety of ego network structures reflects the great diversity of player types. In addition to the obvious variation in the number of partners, we see that some players lie at the center of a squash neighborhood that is very dense (lots of matches between their opponents), and at the other extreme, there are players that seem to be the hub of a group of players, few of whom have played each other. A variety of other highly symmetric formations appear, such as a pentagon circumscribing a star. These are the crop circles of the squash universe. (But lest we become too mystical about such structures, a branch of mathematics known as Ramsey Theory predicts that any sufficiently large random network will reliably produce them.)

Local networks for randomly sampled players

Michael Kearns is a Professor of Computer and Information Science at the University of Pennsylvania and an avid recreational squash player.

Ryan Rayfield is Senior Director of Technology and Strategy at US Squash, and a former undergraduate student of Kearns.

The Erdös Number Project

The Erdös Number Project
Mathematics and Science Center, Room 346
146 Library Drive
Rochester, MI 48309-4479
(**map (/map)**)

Read Aug. 1, 2014 ***News at OU article (//www.oakland.edu/view_news.aspx?sid=34&nid=11554)*** *on the popularity of this website.*

The Erdös Number Project

This is the website for the Erdös Number Project, which studies research collaboration among mathematicians.

The site is maintained by ***Jerry Grossman (https://files.oakland.edu/users/grossman/web/index.html)*** *at* ***Oakland University (//www.oakland.edu)***. ***Patrick Ion****, a retired editor at* ***Mathematical Reviews (http://www.ams.org/publications/60ann/MRPastAndPresent.html)****, and* ***Rodrigo De Castro*** *at the* ***Universidad Nacional de Colombia, Bogota (http://www.matematicas.unal.edu.co/)*** *provided assistance in the past. Please address all comments, additions, and corrections to Jerry at* ***grossman@oakland.edu (mailto:grossman@oakland.edu)***.

Erdös numbers have been a part of the **folklore of mathematicians** throughout the world for many years. For an introduction to our project, a description of what Erdös numbers are, what they can be used for, who cares, and so on, choose the "What's It All About?" link below. To find out who **Paul Erdös (http://www.ams.org/images/erdos-photo-1.gif)** is, look at this **biography (http://www-history.mcs.st-and.ac.uk/history/Mathematicians/Erdos.html)** at the MacTutor History of Mathematics Archive, or choose the "Information about Paul Erdös" link below. Some useful information can also be found in **this Wikipedia article (http://en.wikipedia.org/wiki/Erd%C5%91s_number)**, which may or may not be totally accurate.

WHAT'S INSIDE:

- **What's It All About? (/enp/readme/index)**: General overview, including our (admittedly arbitrary) rules for what counts as a research collaboration.
- **The Data (/enp/thedata/index)**: Lists of all of Paul Erdös's coauthors and their respective coauthors, organized in various ways. There are also links to websites of or about Erdös's coauthors.
- **Facts about Erdös Numbers and Collaborations (/enp/trivia/index)**: Statistical descriptions of Erdös number data, a file of the subgraph induced by Erdös coauthors, Erdös number record holders, facts about collaboration in mathematical research and the collaboration graph, including some information about publishing habits of mathematicians (for example, the median number of papers is 2, and the mean is about 7). This subpage has loads of information about the collaboration graph and Erdös numbers, including the distribution of Erdös numbers (they range up to 13, but the average is less than 5, and almost everyone with a finite Erdös number has a number less than 8) and "Erdös numbers of the second kind".
- **Famous Paths to Paul Erdös (/enp/erdpaths/index)**: Fields Medalists and Nobel Prize winners have small Erdös numbers.
- **Compute Your Own Erdös Number (/enp/compute/index)**: It may be smaller than you think.
- **Research on Collaboration (/enp/research/index)**: Papers on collaboration in scientific research, collaboration graphs and other small world graphs, and Erdös numbers. A lot of research is currently being done by various scientists on collaboration graphs and related topics.
- **Information about Paul Erdös (/enp/erdosdeath/index)**: Information about and links to books, films, articles, memorials, reminiscences.
- **Paul Erdös's Publications (/enp/pubinfo/index)**: They continued to appear more than ten years after his death.
- **Related Concepts (/enp/related/index)**: Six degrees of separation, the Kevin Bacon game, Small Worlds, academic genealogy, Hank Aaron, graph theory.

SPECIAL NOTES:

A nice audio feature (http://relprime.com/erdos) about Paul Erdös and Erdös numbers can be found on the Web.

A segment (//www.npr.org/2013/11/28/207885870/steven-strogatz-the-joy-of-x) of the NPR program "Ask Me Another" featured a quiz about Erdös numbers.

I recommend a delightful children's book about Paul Erdös: The Boy Who Loved Math: The Improbable Life of Paul Erdös (http://deborahheiligman.com/books/the-boy-who-loved-math/).

This website was used as the basis for the 2014 Interdisciplinary Contest in Modeling (http://www.comap.com/undergraduate/contests/mcm/contests/2014/problems/).

There is a project in the works to create a documentary film on Erdös-Bacon numbers (https://www.facebook.com/ErdosBaconMovie/); see also the film's website (http://erdos-bacon.com/).

There is a blog about Erdös numbers (http://www.financial-math.org/blog/2016/11/erdos-numbers-in-finance/) on a site discussing mathematicians on Wall Street.

A new paper by Paul Erdös was published in 2015, with coauthors Ron Graham and Steve Butler. This makes Butler the 512th Erdös coauthor. He and his coauthors will be added to the lists at the next update (around 2020).

NOTES: The data shown on this site are based primarily on all items appearing in **MathSciNet (http://www.ams.org/mathscinet/)** through mid-2015.

If you are an Erdös coauthor, I would really appreciate your sending me (mailto:grossman@oakland.edu) a

complete list of your coauthors (with full names).

One thing we'd really like to do is give more accurate information on some of the old coauthors' status — whether they are still alive. Look at the **list of coauthors arranged by date of first paper with Erdös (https://files.oakland.edu/users/grossman/enp/Erdos0d.html)**; if there is no asterisk after the name, then we assume the person is still alive, except as noted in the **addenda file (/enp/thedata/update/index)**. If anyone has any information that one or more of these are deceased (or, as Paul Erdös would say, "have left"), please **let me know (mailto:grossman@oakland.edu)**. (We know some are alive; please report only those that have passed on, and report only Erdös coauthors, since there is no way we could extend this convention to those with Erdös number 2.)

I have deleted all links to the unsolicited translations that used to be here, because in some cases people were providing poor translations and linking to commercial sites.

URL = http://www.oakland.edu/enp
This page was last updated on July 6, 2017 (but subpages may have been updated more recently).
However, the lists of coauthors and the various other statistics on this site are updated about once every five years. The current version was posted on July 14, 2015 and includes all information listed in **MathSciNet (http://www.ams.org/mathscinet/)** and **DBLP (http://dblp.uni-trier.de/db/index.html)** through mid-2015.

Business Administration (/business)

Education and Human Services (/sehs)

Engineering and Computer Science (/secs)

Health Sciences (/shs)

Nursing (/nursing)

OUWB School of Medicine (/medicine)

Graduate Education (/grad)

Honors College (/hc)

Integrative Studies (/bis)

INFO FOR

Future Undergraduate Students (/futurestudents)

Future Graduate Students (/grad)

Current Students (/students)

Alumni (http://www.oualumni.com/s/1001/1001-alumni/start.aspx?gid=1001&pgid=1117)

Faculty and Staff (/faculty-and-staff)

Donors (http://www.isupportou.com/s/1001/04-giving/start.aspx?gid=4&pgid=61)

QUICK LINKS

About OU (/about)

Important Dates (/registrar/important-dates/)

Diversity, Equity and Inclusion (/diversity/)

Directory (/directory)

Jobs at OU (https://jobs.oakland.edu)

University Offices (/universityoffices)

Webmaster (mailto:webmaster@oakland.edu)

OU INC (/ouinc)

Macomb-OU Incubator (/macombouic)

Macomb Programs (/macomb/)

Eye Research Institute (/eri)

Can cascades be predicted?

Justin Cheng
Stanford University
jcccf@cs.stanford.edu

Lada A. Adamic
Facebook
ladamic@fb.com

P. Alex Dow
Facebook
adow@fb.com

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

ABSTRACT

On many social networking web sites such as Facebook and Twitter, resharing or reposting functionality allows users to share others' content with their own friends or followers. As content is reshared from user to user, large cascades of reshares can form. While a growing body of research has focused on analyzing and characterizing such cascades, a recent, parallel line of work has argued that the future trajectory of a cascade may be inherently unpredictable. In this work, we develop a framework for addressing cascade prediction problems. On a large sample of photo reshare cascades on Facebook, we find strong performance in predicting whether a cascade will continue to grow in the future. We find that the relative growth of a cascade becomes more predictable as we observe more of its reshares, that temporal and structural features are key predictors of cascade size, and that initially, breadth, rather than depth in a cascade is a better indicator of larger cascades. This prediction performance is robust in the sense that multiple distinct classes of features all achieve similar performance. We also discover that temporal features are predictive of a cascade's eventual shape. Observing independent cascades of the same content, we find that while these cascades differ greatly in size, we are still able to predict which ends up the largest.

Categories and Subject Descriptors: H.2.8 **[Database Management]**: Database applications—*Data mining*
General Terms: Experimentation, Measurement.
Keywords: Information diffusion, cascade prediction, contagion.

1. INTRODUCTION

The sharing of content through social networks has become an important mechanism by which people discover and consume information online. In certain instances, a photo, link, or other piece of information may get *reshared* multiple times: a user shares the content with her set of friends, several of these friends share it with their respective sets of friends, and a *cascade* of resharing can develop, potentially reaching a large number of people. Such cascades have been identified in settings including blogging [1, 13, 21], e-mail [12, 22], product recommendation [20], and social sites such as Facebook and Twitter [9, 18]. A growing body of research has focused on characterizing cascades in these domains, including their structural properties and their content.

In parallel to these investigations, there has been a recent line of work adding notes of caution to the study of cascades. These cautionary notes fall into two main genres: first, that large cascades are rare [11]; and second, that the eventual scope of a cascade may be an inherently unpredictable property [28, 31]. The first concern — that large cascades are rare — is a widespread property that has been observed quantitatively in many systems where information is shared. The second concern is arguably more striking, but also much harder to verify quantitatively: to what extent is the future trajectory of a cascade predictable; and which features, if any, are most useful for this prediction task?

Part of the challenge in approaching this prediction question is that the most direct ways of formulating it do not fully address the two concerns above. Specifically, if we are presented with a short initial portion of a cascade and asked to estimate its final size, then we are faced with a pathological prediction task, since almost all cascades are small. Alternately, if we radically overrepresent large cascades in our sample, we end up studying an artificial setting that does not resemble how cascades are encountered in practice. A set of recent initial studies have undertaken versions of cascade prediction despite these difficulties [19, 23, 26, 29], but to some extent they are inherent in these problem formulations.

These challenges reinforce the fact that finding a robust way to formulate the problem of cascade prediction remains an open problem. And because it is open, we are missing a way to obtain a deeper, more fundamental understanding of the predictability of cascades. How should we set up the question so that it becomes possible to address these issues directly, and engage more deeply with arguments about whether cascades might, in the end, be inherently unpredictable?

The present work: Cascade growth prediction. In this paper, we propose a new approach to the prediction of cascades, and show that it leads to strong and robust prediction results. We are motivated by a view of cascades as complex dynamic objects that pass through successive stages as they grow. Rather than thinking of a cascade as something whose final endpoint should be predicted from its initial conditions, we think of it as something that should be *tracked* over time, via a sequence of prediction problems in which we are constantly seeking to estimate the cascade's next stage from its current one.

What would it mean to predict the "next stage" of a cascade? If we think about all cascades that reach size k, there is a distribution of eventual sizes that these cascades will reach. Then the distribution of cascade sizes has a median value $f(k) \geq k$. This number $f(k)$ is thus the "typical" final size for cascades that reached size

WWW '14 Seoul, Republic of Korea
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2567997.

at least k. Hence, the most basic way to ask about a cascade's next stage of growth, given that it currently has size k, is to ask whether it reaches size $f(k)$.

We therefore propose the following *cascade growth prediction problem*: given a cascade that currently has size k, predict whether it grow beyond the median size $f(k)$. (As we show later, the prediction problem is equivalent to asking: given a cascade of size k, will the cascade double its size and reach at least $2k$ nodes?) This implicitly defines a family of prediction problems, one for each k. We can thus ask how cascade predictability behaves as we sweep over larger and larger values of k. (There are natural variants and generalizations in which we ask about reaching target sizes other than the median $f(k)$.) This problem formulation has a number of strong advantages over standard ways of trying to define cascade prediction. First, it leads to a prediction problem in which the classes are balanced, rather than highly unbalanced. Second, it allows us to ask for the first time how the predictability of a cascade varies over the range of its growth from small to large. Finally, it more closely approximates the real tasks that need to be solved in applications for managing viral content, where many evolving cascades are being monitored, and the question is which are likely to grow significantly as time moves forward.

For studying cascade growth prediction, it is important to work with a system in which the sharing and resharing of information is widespread, the complete trajectories of many cascades—both large and small—are observable, and the same piece of content shared separately by many people, so that we can begin to control for variation in content. For this purpose, we use a month of complete photo-resharing data from Facebook, which provides a rich ecosystem of shared content exhibiting all of these properties.

In this setting, we focus on several categories of questions:

(i) How high an accuracy can we achieve for cascade growth prediction? If we cannot improve on baseline guessing, then this would be evidence for the inherent unpredictably of cascades. But if we can significantly improve on this baseline, then there is a basis for non-trivial prediction. In the latter case, it also becomes important to understand the features that make prediction possible.

(ii) Is growth prediction more tractable on small cascades or large ones? In other words, does the future behavior of a cascade become more or less predictable as the cascade unfolds?

(iii) Beyond just the growth of a cascade, can we predict its "shape" — that is, its network structure?

Summary of results. Given the challenges in predicting cascades, we find surprisingly strong performance for the growth prediction problem. Moreover, the performance is robust in the sense that multiple distinct classes of features, including those based on time, graph structure, and properties of the individuals resharing, can achieve accuracies well above the baseline. Cascades whose initial reshares come quickly are more likely to grow significantly; and from a structural point of view, breadth rather than depth in the resharing tree is a better predictor of significant growth.

We investigate the performance of growth prediction as a function of the size of the cascade so far — when we want to predict the growth of a cascade of size k, how does our accuracy depend on k? It is not a priori clear whether accuracy should increase or decrease as a function of k, since for any value of k the challenge is to determine what the cascade will do in the future. Seeing more of the cascade (larger k) does not make the problem easier, as it also involves predicting "farther" into the future (i.e., whether the cascade will reach size at least $2k$). We find that accuracy increases with k, so that it is possible to achieve better performance on large cascades than small ones. The features that are most significant for prediction change with k as well, with properties of the content and the original author becoming less important, and temporal features remaining relatively stable.

We also consider a related question: how much of a cascade do we need to see in order to obtain good performance? Specifically, suppose we want to predict the growth of a cascade of size at least R, but we are only able to see the first $k < R$ nodes in the cascade. How does prediction performance depend on k, and in particular, is there a "sweet spot" where a relatively small value of k gives most of the performance benefits? We find in fact that there is no sweet spot: performance essentially climbs linearly in k, all the way up to $k = R$. Perhaps surprisingly, more information about the cascade continues to be useful even up to the full snapshot of size R.

In addition to growth, we also study how well we can predict the eventual "shape" of the cascade, using metrics for evaluating tree structures as a numerical measure of the shape. We obtain performance significantly above baseline for this task as well; and perhaps surprisingly, multiple classes of features including temporal ones perform well for this task, despite the fact that the quantity being predicted is a purely structural one.

One of the compelling arguments that originally brought the issue of inherent unpredictability onto the research agenda was a striking experiment by Salganik, Dodds, and Watts, in which they showed that the same piece of content could achieve very different levels of popularity in separate independent settings [28]. Given the richness of our data, we can study a version of this issue here in which we can control for the content being shared by analyzing many cascades all arising from the sharing of the same photo. As in the experiment of Salganik et al., we find that independent resharings of the same photo can generate cascades of very different sizes. But we also show that this observation can be compatible with prediction: after observing small initial portions of these distinct cascades for the same photo, we are able to predict with strong performance which of the cascades will end up being the largest. In other words, our data shows wide variation in cascades for the same content, but also predictability despite this variation.

Overall, our goal is to set up a framework in which prediction questions for cascades can be carefully analyzed, and our results indicate that there is in fact a rich set of questions here, pointing to important distinctions between different types of features characterizing cascades, and between the essential properties of large and small cascades.

2. RELATED WORK

Many papers have analyzed and cataloged properties of empirically observed information cascades, while others have considered theoretical models of cascade formation in networks. Most relevant to our work are those which focus on predicting the future popularity of a given piece of content. These studies have proposed rich sets of features for prediction, which we discuss later in Section 3.2.

Much prior work aims to predict the *volume* of *aggregate* activity — the total number of up-votes on Digg stories [29], total hourly volume of news phrases [34], or total daily hashtag use [23]. At the other end of the spectrum, research has focused on *individual* user-level prediction tasks: whether a user will retweet a specific tweet [26] or share a specific URL [10]. Rather than attempt to predict aggregate popularity or individual behavior in the next time step, we instead look at whether an information cascade grows over the median size (or doubles in size, as we later show).

Research on communities defined by user interests [3] or hashtag content [27] has also looked at a notion of growth, predicting

whether a group will increase in size by a given amount. Nevertheless, these focused on groups of already non-trivial size, and their growth predicted without an explicit internal cascade topology, and without tracking predictability over different size classes.

Several papers focus on predictions after having observed a cascade for a given fixed time frame [19, 23, 30]. In contrast, rather than studying specific time slices, we continuously observe the cascade over its entire lifetime and attempt to understand how predictive performance varies as the cascade develops. Moreover, our methodology does not penalize slowly but persistently growing cascades. Thus, we predict the size and the structure after having observed a certain number of initial reshares.

Many studies consider the cascade prediction task as a regression problem [6, 19, 29, 30] or a binary classification problem with large bucket sizes [16, 17, 19]. The danger with these approaches is that they are biased towards studying extremely large but also extremely rare cascades, bypassing the whole issue about the general predictability of cascades. For example, research has specifically focused on content and users that create extremely large cascades, such as popular hashtags [15, 33] and very popular users [9, 14], which has led to criticism that cascades may only be predictable after they have already grown large [31]. While it is useful to understand the dynamics of extremely popular content, such content is also very rare. Thus, we rather seek to understand predictability along cascade's entire lifetime. We consider cascades that have as few as five reshares, and introduce a classification task which is not skewed towards very large cascades.

3. PREDICTING CASCADE GROWTH

To examine the cascade growth prediction problem, we first define and motivate our experimental setup and the feature sets used, then report our prediction results with respect to different k.

3.1 Experimental setup

Mechanics of information passing on Facebook. We focus on content consisting of posts the author has designated as public, meaning that anyone on Facebook is eligible to view it, and we further restrict our attention to content in the form of photos, which comprise the majority of reshare cascades on Facebook [9]. Such posts are then distributed by Facebook's News Feed, typically at first to users who are either friends of the poster or who subscribe to their content, e.g. as followers. Each post is accompanied by a "share" link that allows friends and followers to "reshare" the post with her own friends and followers, thus expanding the set of users exposed to the content. This explicit sharing mechanism creates information cascades, starting with the root node (user or page) that originally created the content, and consisting of all subsequent reshares of that content.

Figure 1 illustrates the process with an example: a node v_0 posts a public photo, seen by v_0's friends and followers in their News Feeds. Friends v_1 and v_3 then share the photo with their own friends. This way the photo propagates over the edges of the Facebook network and creates an information cascade. We represent the cascade graph as $\hat{G}$, and the induced subgraph of all photo sharers, including all friendship or follow links between them as G'. Notice that some users (ex. v_5) are exposed via multiple sources (v_0, v_1, v_3, v_4).

An important issue for our understanding of reshare cascades is the following distinction: content can be produced by *users* — individual Facebook accounts whose primary audience consists of friends and any subscribers the individual has — and it can also be produced by *pages*, which correspond to the Facebook accounts of companies, brands, celebrities, and other highly visible public entities. In the common parlance around cascades, reshared content originally produced by a user is often informally viewed as more "organic," developing a following in a more bottom-up way. In contrast, reshared content from pages is seen as more top-down, and generally broadcast via News Feed to a larger set of initial followers. A natural question, and a theme that will run through several analyses in the paper, is to understand if these distinctions carry over to the properties we study here: do user-initiated cascades differ in their predictability and their underlying structure from page-initiated cascades?

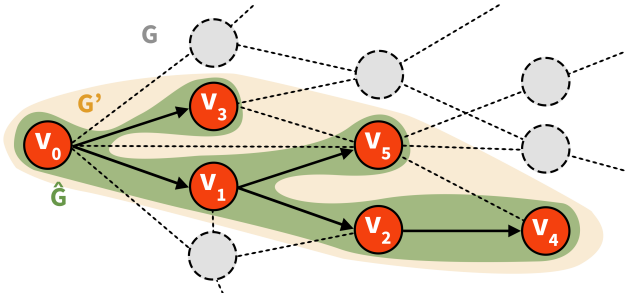


Figure 1: An information cascade represented by solid edges on a graph G, starting at v_0 ($\hat{G}$). Dashed lines indicate friendship edges; the edges between resharers make up the friend subgraph G'.

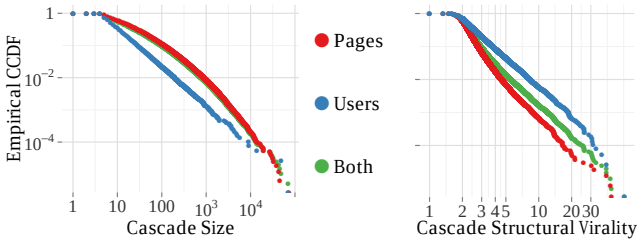
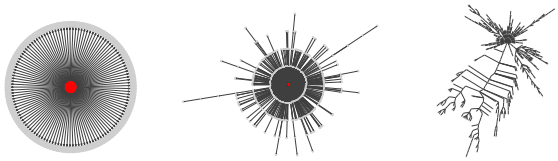


Figure 2: The complementary cumulative distribution (CCDF) of cascade size (left) and structural virality measured by using the Wiener index (right).

Dataset description. We sampled our anonymized dataset from photos uploaded to Facebook in June 2013 and observed any reshares occurring within 28 days of initial upload. The dataset only includes photos posted publicly (viewable by anyone), and not deleted during the observation period. Further, we exclude photos with fewer than five reshares as is required by the prediction tasks described below. We constructed diffusion trees first by taking the explicit cascade, e.g. C clicking "share" on B's "share" of A's photo forms the cascade $A \to B \to C$. However, it is possible that user C clicked on user B's share, and then directly reshared from A. Since we want to know how the information actually flowed in the network, we reconstruct the path $A \to B \to C$ based on click, impression, and friend/follower data [9].

Figure 2 begins to show how photos uploaded by pages generate cascades that differ from those uploaded by users. In our dataset, 81% of cascades are initiated by pages. Figure 2 shows the cascade size distribution for pages, users, and the two combined. Page cascades are typically larger than user cascades, e.g., 11% of page cascades reach at least 100 reshares, while only 2% of user cascades do, though both follow heavy tailed distributions. Fitting power-law curves to their tails, we observe power-law exponents of α equal to 2.2, 2.1, and 2.1 for user, page, and both, respectively ($x_{\min} = 10, 2000, 2000$).

In addition to cascade growth, we quantify the shape of a cascade using the Wiener index, defined as the average distance between all



(a) $d = 1.98$ (b) $d = 2.47$ (c) $d = 14.4$

Figure 3: Cascades with a low Wiener index d resemble star graphs, while those with a high index appear more viral (the root is red).

pairs of nodes in a cascade. Recent work has proposed the Wiener index as a measure of the structural virality of a cascade [2]. Figure 3 shows examples of cascades with varying Wiener index values. Intuitively, a cascade with low structural virality has most of its distribution following from a small number of hub nodes, while a cascade with high virality will have many long paths. Figure 2 shows the distribution of cascade virality (as measured by Wiener index) in our dataset, which, as we saw with cascade size, follows a heavy-tailed distribution. While user cascades are typically smaller than page cascades in our dataset, they tend to have greater structural virality, supporting the intuition that the structure of user-initiated cascades is richer and deeper than that of page-initiated cascades.

Defining the cascade growth prediction problem. Our aim in this paper is to study how well cascades can be predicted. Moreover, we are interested in understanding how various aspects of the prediction task affect the predictive performance.

There are several formulations of the task. If we were to define the task as a regression problem, predictions may be skewed towards large cascades, as cascade size follows a heavy-tailed distribution (Figure 2(right)). Similarly, if we define it as a classification problem of predicting whether a cascade reaches a specific size, we may end up with unbalanced classes, and an overrepresentation of large cascades. Also, if we simply observed a small initial portion of a cascade, and predict its future size, the problem is pathological as almost all cascades are small. And, if we only varied the initial period of observation, the task of predicting whether a cascade reaches a certain size gets easier as we observe more of it.

To remedy these issues, we define a classification task that does not suffer from these deficiencies. We consider a binary classification problem where we observe the first k reshares of a cascade and predict whether the eventual size of a cascade reaches the median size of all the cascades with at least k reshares, $f(k)$. This allows us to study how cascade predictibility varies with k. As exactly half the cascades reach a size greater than the median by definition, random guessing achieves accuracy of 50%.

Interestingly, the question of whether the cascade will reach $f(k)$ is equivalent to that of whether a cascade will double in size. This follows directly from the fact that cascade size distribution follows a power-law with exponent $\alpha \approx 2$. Consider a power-law distribution on the interval (x_{min}, ∞) with a power-law exponent $\alpha \approx 2$. Then the median $f(x)$ of this distribution is $2 \cdot x_{min}$, as demonstrated by the following calculation:

$$\int_{x_{\min}}^{f(x)} \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} dx = \frac{1}{2} \Rightarrow f(x) = 2^{\frac{1}{\alpha - 1}} x_{\min} = 2x_{\min}$$

As we examine cascades of size greater than $k = x_{min}$, the median size of these cascades is thus $2 \cdot k$ from this derivation. In each of our prediction tasks, we observe that this is indeed true.

Methods used for learning. Our general methodology for the cascade prediction problem will be to represent a cascade by a set of features and then use machine learning classifiers to predict its future size. We used a variety of learning methods, including linear regression, naive Bayes, SVM, decision trees and random forests. However, we primarily report performance of the logistic regression classifier for ease of comparison. In many cases, the performance of most classifiers was similar, although non-linear classifiers such as random forests usually performed slightly better than linear classifiers such as logistic regression. In all cases, we performed 10-fold cross validation and report the classification accuracy, F1 score, and area under the ROC curve (AUC).

3.2 Factors driving cascade growth

We proceed by describing factors that contribute to the growth and spreading of cascades. We group these factors into five classes: properties of the content that is spreading, features of the original poster, features of the resharer, structural features of the cascade, and temporal characteristics of the cascade. Table 1 contains a detailed list of features.

Content features. The first natural factor contributing to the ability of the cascade to spread is the content itself [7]. On Twitter, tweet content and in particular, hashtags, are used to generate content features [23, 30], and identify topics affecting retweet likelihood [26]. LDA topic models have also been incorporated into these prediction tasks [16], and human raters employed to infer the interestingness of content [5, 26]. In our work, we relied on a linear SVM model, trained using image GIST descriptors and color histogram features, to assign likelihood scores of a photo being a closeup shot, taken indoors or outdoors, synthetically generated (e.g., screenshots or pure text vs. photographs), or contained food, a landmark, person, nature, water, or overlaid text (e.g., a meme). We also analyzed words in the caption accompanying an image for positive sentiment, negative sentiment, and sociality [17, 25].

Nevertheless, while content features affected the performance of structural and temporal features, we find that they are weak predictors of how widely disseminated a piece of content would become.

Original poster/resharer features. Some prior work focused on features of the root note in a cascade to predicting the cascade's evolution, finding that content from highly-connected individuals reaches larger audiences, and thus spreads further. Users with large follower counts on Twitter generated the largest retweet cascades [5]. Separately, features of an author of a tweet were shown to be more important than features of the tweet itself [26]. In many Twitter studies predicting cascade size or popularity, a user's number of followers ranks among the top, if not the most, important predictor of popularity [5, 23].

Other features of the root node have also been studied, such as the number of prior retweets of a user's posts [5, 16], and how many Twitter lists a user was included in [26]. The number of @-mentions of a Twitter user was used to predict whether, and how soon a tweet would be retweeted, how many users would directly retweet, and the depth a cascade would reach [33]. Still, [8] found that various measures of a user's popularity are not very correlated with his or her influence.

We capture the intuition behind these factors by defining demographic as well as network features of the original poster as well as the features of the users who reshared the content so far. We use Facebook's distinction of users (individuals) and pages (entities representing an interest) to further distinguish different origin types, in addition to the influence features mentioned above.

Structural features of the cascade. Networks provide the substrate through which information spreads, and thus their structure influences the path and reach of the cascade. As illustrated in Fig-

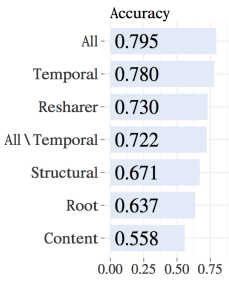


Figure 4: Using logistic regression, we are able to predict with near 80% accuracy whether the size of a cascade will reach the median (10) after observing the first k =5 reshares.

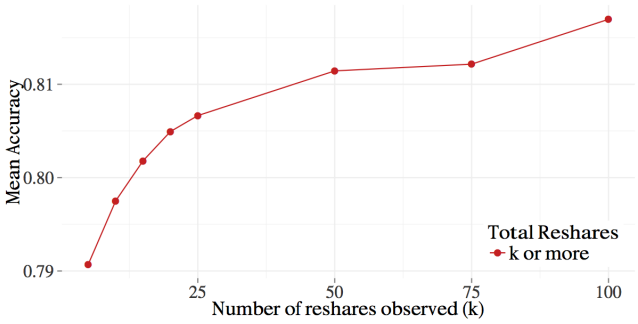


Figure 5: If we observe the first k reshares of a cascade, and want to predict whether the cascade will double in size, our prediction improves as we observe more of it.

ure 1, we generate features from both the graph of the first k reshares ($\hat{G}$), as well as the induced friend subgraph of the first k resharers (G'). Whereas the reshare graph $\hat{G}$ describes the actual spread of a cascade, the friend subgraph G' provides information about the social ties between these initial resharers. The social graph G allows us to compute the potential reach of these reshares.

Previous work considered the network structure of the underlying graph in inferring the virality of content [32], with highly viral items spreading across communities. We use the density of the initial reshare cascade ($subgraph'_k$) and the proximity to the root node ($orig\_connections_k$, did_leave) as proxies for whether an item is spreading primarily within a community or across many. One can also look outside the network between resharers, and count the number of users reachable via all friendship and follow edges of the first k users ($border\_nodes_k$). This relates to total number of exposed users, and has been demonstrated to be an important feature in predicting Twitter hashtag popularity [23].

As we can trace information flow on Facebook exactly, we need not worry about independent entry points influencing a cascade [6, 24]; external influence instead allows us to investigate multiple independent cascades arising from the same content (see Section 5.1).

Temporal features. Properties related to the "speed" of the cascade (e.g., $time_k$) were shown to be the most important features in predicting thread length on Facebook [4], and are a primary mechanism in predicting online content popularity [29]. Moreover, as the speed of diffusion changes over time, this may have a strong effect on the ability of the cascade to continue spreading through the network [33].

We characterize a number of temporal properties of cascade diffusion (see Table 1). In particular, we measure the change in the speed of reshares ($time''_{1..k}$), compare the differences between the speed in the first and second half of the measurement period ($time'_{1..k/2}$, $time'_{k/2..k}$), and qunatify the number of users who were exposed to the cascade per time unit ($views'_{1..k-1,\ k}$).

3.3 Predicting cascade growth

To illustrate the general performance of the features described in the previous section we consider a simple prediction task, where we observe the first 5 reshares of the cascade and want to predict whether it will reach the median cascade size (or equivalently, whether it will double and be reshared at least 10 times). For the experiment we use a set of N_c = 150,572 photos, where each photo was shared at least 5 times. The total number of reshares of these photos was N_r = 9,233,300.

Figure 4 shows logistic regression performance using all features from Table 1. For this task, random guessing would obtain a performance of 0.5, while our method achieves surprisingly strong performance: classification accuracy of 0.795 and AUC of 0.877. If we relax the task and instead of predicting above vs. below median size, we predict top vs. bottom quartile (top 25% vs. bottom 25%) the accuracy rises even further to 0.926, and the AUC to 0.976.

Overall, while each feature set is individually significantly better than predicting at random, it is the set of temporal features that outperforms all other individual feature sets, obtaining performance scores within 0.025 of those obtained when using all features. To understand if we could do well without temporal features, we trained a classifier which excluded them and were still able to obtain reasonable performance even without these features. This is especially useful when one knows through *whom* information was passed, but not *when* it was passed. The lack of reliance on any individual set of features demonstrates that the predictions are robust.

Studied individually, we also find that temporal features generally performed best, followed by structural features. The reshare rate in the second half ($time'_{k/2..k}$) was most predictive, attaining accuracy of 0.73. This was followed by the rate of user views of the original photo, $views'_{0,k}$, and the time elapsed between the original post and fifth reshare, $time_5$ (both 0.72). In fact, $time_{k+1}$ is always more accurate than $time_k$. The most accurate structural features were did_leave and $outdeg(v_0)$ (both 0.65). We examine individual feature importance in more detail later.

3.4 Predictability and the observation window of size k

It is also natural to ask whether cascades get more or less predictable as we observe more of the initial growth of a cascade. One may think that observing more of the cascade may allow us to extrapolate its future growth better; on the other hand, additional observed reshares may also introduce noise and uncertainty in the future growth of the cascade. Note that the task does not get easier as we observe more of the cascade, as we are predicting whether the cascade will reach size $2k$ (or equivalently, the median) given that we have seen k reshares so far.

Figure 5 shows that the predictive performance of whether a cascade doubles in size increases as a function of the number of observed reshares k. In other words, it is easier to predict whether a cascade that has reached 25 reshares will get another 25, than to predict whether a cascade that has reached 5 reshares will obtain an additional 5. Thus, the prediction accuracy for larger cascades is above the already high accuracy for smaller values of k. The change in the F1 score and AUC also follow a very similar trend.

Overall, these results demonstrate that observing more of the cascade, while also predicting "farther" into the future, is easier than observing a cascade early in its life and predicting what it will

	Content Features
$score_{food/nature/...}$	The probability of the photo having a specific feature (food, overlaid text, landmark, nature, etc.)
is_en	Whether the photo was posted by an English-speaking user or page
$has\_caption$	Whether the photo was posted with a caption
$liwc_{pos/neg/soc}$	Proportion of words in the caption that expressed positive or negative emotion, or sociality, if English
	Root (Original Poster) Features
$views_{0,\,k}$	Number of users who saw the original photo until the kth reshare was posted
$orig\_is\_page$	Whether the original poster is a page
$outdeg(v_0)$	Friend, subscriber or fan count of the original poster
age_0	Age of the original poster, if a user
$gender_0$	Gender of the original poster, if a user
fb_age_0	Time since the original poster registered on Facebook, if a user
$activity_0$	Average number of days the original poster was active in the past month, if a user
	Resharer Features
$views_{1..k-1,\,k}$	Number of users who saw the first $k-1$ reshares until the kth reshare was posted
$pages_k$	Number of pages responsible for the first k reshares, including the root, or $\sum_{i=0}^{k} \mathbb{1}\{v_i \text{ is a page}\}$
$friends_k^{avg/90p}$	Average or 90th percentile friend count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{friends}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fans_k^{avg/90p}$	Average or 90th percentile fan count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg(v_i)\mathbb{1}\{v_i \text{ is a page}\}$
$subscribers_k^{avg/90p}$	Average or 90th percentile subscriber count of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} outdeg_{subscriber}(v_i)\mathbb{1}\{v_i \text{ is a user}\}$
$fb\_ages_k^{avg/90p}$	Average or 90th percentile time since the first k resharers registered on Facebook, or $\frac{1}{k}\sum_{i=1}^{k} fb\_age_i$
$activities_k^{avg/90p}$	Average number of days the first k resharers were active in July, or $\frac{1}{k}\sum_{i=1}^{k} activity_i$
$ages_k^{avg/90p}$	Average age of the first k resharers, or $\frac{1}{k}\sum_{i=1}^{k} age_i$
$female_k$	Number of female users among the first k resharers, or $\sum_{i=1}^{k} \mathbb{1}\{gender_i \text{ is female}\}$
	Structural Features
$outdeg(v_i)$	Connection count (sum of friend, subscriber and fan counts) of the ith resharer (or out-degree of v_i on $G=(V,E)$)
$outdeg(v_i')$	Out-degree of the ith reshare on the induced subgraph $G'=(V',E')$ of the first k resharers and the root
$outdeg(\hat{v}_i)$	Out-degree of the ith reshare on the reshare graph $\hat{G}=(\hat{V},\hat{E})$ of the first k reshares
$orig\_connections_k$	Number of first k resharers who are friends with, or fans of the root, or $\lvert\{v_i \mid (v_0,v_i)\in E,\ 1\le i\le k\}\rvert$
$border\_nodes_k$	Total number of users or pages reachable from the first k resharers and the root, or $\lvert\{v_i \mid (v_i,v_j)\in E,\ 0\le i,j\le k\}\rvert$
$border\_edges_k$	Total number of first-degree connections of the first k resharers and the root, or $\lvert\{(v_i,v_j) \mid (v_i,v_j)\in E,\ 0\le i,j\le k\}\rvert$
$subgraph_k'$	Number of edges on the induced subgraph of the first k resharers and the root, or $\lvert\{(v_i,v_j) \mid (v_i,v_j)\in E',\ 0\le i,j\le k\}\rvert$
$depth_k'$	Change in tree depth of the first k reshares, or $\min_\beta \sum_{i=1}^{k}(depth_i-\beta i)^2$
$depths_k^{avg/90p}$	Average or 90th percentile tree depth of the first k reshares, or $\frac{1}{k}\sum_{i=1}^{k} depth_i$
did_leave	Whether any of the first k reshares are not first-degree connections of the root
	Temporal Features
$time_i$	Time elapsed between the original post and the ith reshare
$time'_{1..k/2}$	Average time between reshares, for the first $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=1}^{k/2-1}(time_{i+1}-time_i)$
$time'_{k/2..k}$	Average time between reshares, for the last $k/2$ reshares, or $\frac{1}{k/2-1}\sum_{i=k/2}^{k-1}(time_{i+1}-time_i)$
$time''_{1..k}$	Change in the time between reshares of the first k reshares, or $\min_\beta \sum_{i=1}^{k-1}(time_{i+1}-time_i)-\beta i)^2$
$views'_{0,k}$	Number of users who saw the original photo, until the kth reshare was posted, per unit time, or $\frac{views_{0,\,k}}{time_k}$
$views'_{1..k-1,\,k}$	Number of users who saw the first $k-1$ reshares, until the kth reshare was posted, per unit time, or $\frac{views_{1..k-1,k}}{time_k}$

Table 1: List of features used for learning. We compute these features given the cascade until the kth reshare.

do next (i.e., $k=5$ vs. $k=25$).

Fixing the minimum cascade size R. In the previous version of the task, cascades are required only to have at least k reshares. Thus, the set of cascades changes with k. Here, we examine a variation of this task, where we compose a dataset of cascades that have at least R reshares. We observe the first k ($k \le R$) reshares of the cascade and aim to predict whether the cascade will grow over the median size (over all cascades of size $\ge R$). As we increase k, the task gets easier as we observe more of the cascade and the predicted quantity does not change.

With the task, we find that performance increases linearly with k up to R, or that there is no "sweet spot" or region of diminishing returns ($p < 0.05$ using a Harvey-Collier test). For example, the top-most line in Figure 6 shows that when each observed cascade has obtained 100 or more reshares, performance increases linearly as more of the cascade is observed. This demonstrates that more information is always better: the greater the number of observed reshares, the better the prediction.

However, Figure 6 also shows that larger cascades are less predictable than smaller cascades. For example, predicting whether cascades with 1,000 to 2,000 reshares grow large is significantly more difficult than predicting cascades of 100 to 200 reshares. This

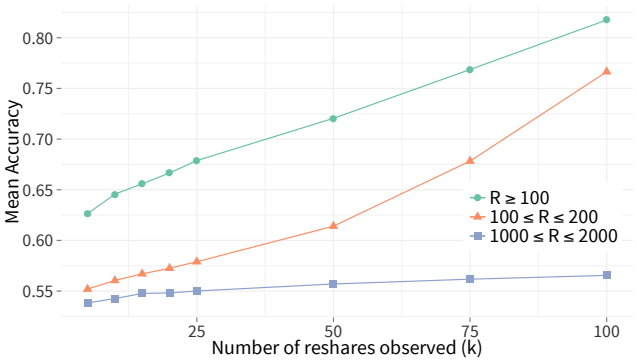


Figure 6: Knowing that a cascade obtains at least R reshares, prediction performance increases linearly with k, $k \leq R$. However, differentiating among cascades with large R also becomes more difficult.

shows that once one knows that a cascade will grow to be large, knowing the characteristics of the very beginning of its spread is less useful for prediction.

3.5 Changes in feature importance

We now examine how feature importance changes as more and more of the cascade is observed. In this experiment, we compute the value of the feature after observing first k reshares and measure the correlation coefficient of the feature value with the log-transformed number of reshares (or cascade size).

Figure 7 shows the results for the five feature types. We summarize the results by the following observations:

- *Correlations of averages increase with the number of observations*. As we obtain more examples, naturally averages get less noisy, and more predictive (e.g., $ages^{avg}$ and $friends^{avg}$).
- *The original post gets less important with increasing k*. After observing 100 reshares, it becomes less important that the original post was made by a page ($orig\_is\_page$), or that the original poster had many connections to other users ($outdeg(v_0)$).
- *Similarly, the actual content being reshared gets less important with increasing k*. Almost all content features tend to zero as k increases, except for $has\_caption$ and is_en. This can be explained by the fact that cascades of photos with captions have a unimodal distribution, and cascades started by English speakers have a bimodal distribution. Thus, these features become correlated in opposite directions.
- *Successful cascades get many views in a short amount of time, and achieve high conversion rates*. The number of users who have viewed reshares of a cascade is more negatively correlated with increasing k ($views_{1..k-1,k}$), suggesting that requiring "fewer tries" to achieve a given number of reshares is a positive indicator of its future success. On the other hand, while requiring fewer views is good, rapid exposure, or reaching many users within a short amount of time is also a positive predictor ($views'_{1..k-1,k}$).
- *Structural connectedness is important, but gets less important over time*. Nevertheless, reshare depth remains highly correlated: the deeper a cascade goes, the more likely it is to be long-lasting, as even users "far away" from the original poster still find the content interesting.

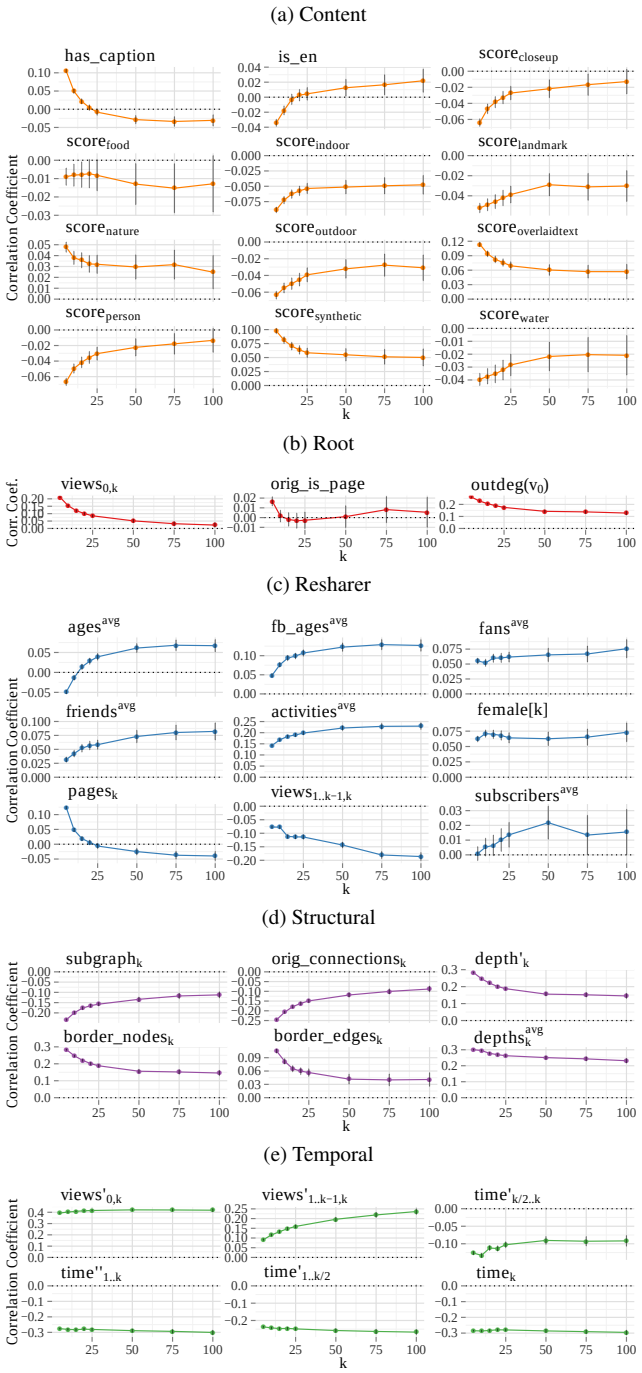


Figure 7: The importance of each feature varies as we observe more of a cascade, as shown by the change in correlation coefficients.

- *The importance of timing features remains relatively stable*. While highly correlated, timing features remain remarkably stable in importance as k increases.

We note individual features' logistic regression coefficients empirically follow similar shapes, but have the downside of having interactions with one another. Using either the slope of the best-fit line of the cascade size against the normalized feature value, or individual feature performance also reveals similar trends. Further LIWC text content features (positive, negative, and social categories) consistently performed poorly, attaining performance no

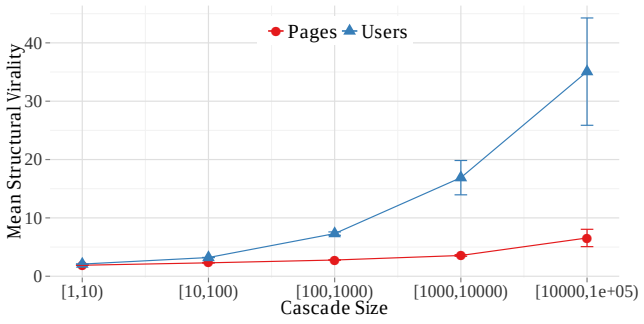


Figure 8: The mean structural virality (Wiener index) increases with cascade size, but is significantly higher for user cascades.

better than chance, with accuracy between 0.49 and 0.52.

4. PREDICTING CASCADE STRUCTURE

Similar to predicting cascade size we can also attempt to predict the *structure* of the cascade. We now turn to examining how structural features of the cascade determine its evolution and spread.

4.1 User-started and page-started cascades

Earlier we discussed the notion of *structural virality* as a measure of how much the structure of a cascade is dominated by a few hub nodes, and we saw that user-initiated cascades have significantly higher structural virality than page-initiated cascades, reflecting their richer graph structure. It is natural to ask how these distinctions vary with the size of the cascade — are large user-initiated cascades more similar to page-initiated ones, e.g. are they driven by popular hub nodes?

Figure 8 shows that the opposite is the case — user and page-initiated cascades remain structurally distinct, with this distinction even increasing with cascade size. Moreover, this difference continues to hold even when controlling for the number of first-degree reshares (directly from the root), suggesting a certain robustness to their richer structure. Because of these structural differences, we handle user and page cascades separately in the analyses that follow.

These distinctions may also help explain a large difference in the predictability of user-initiated vs. page-initiated cascades. We observe that for page cascades accuracy exceeds 80%, while that for user cascades is slightly under 70%. (These results also hold for the F1 score and AUC, with a difference of about 0.1.) The fact that much more of the structure of a page-initiated cascade is typically carried by a small number of hub nodes may suggest why the prediction task is more tractable in this case.

4.2 The initial structure of a cascade influences its eventual size

To understand how structure bears on the future growth of the cascade, we examine how the configuration of the first three reshares (and the root) correlates with the cascade size. In particular, we measure the proportion of cascades starting from each configuration that reach the median size. We do this separately for two different initial poster types: a user, and a page. We discard "celebrity" users who may large followings like the most popular pages. Figure 9a shows that as the initial cascade structure becomes shallower, the proportion of cascades that double in size increases. To examine why this would be the case, we also examined the time needed for the 3rd reshare to happen (Figure 9c). For pages, shallower cascades tend to happen more rapidly, consistent with being

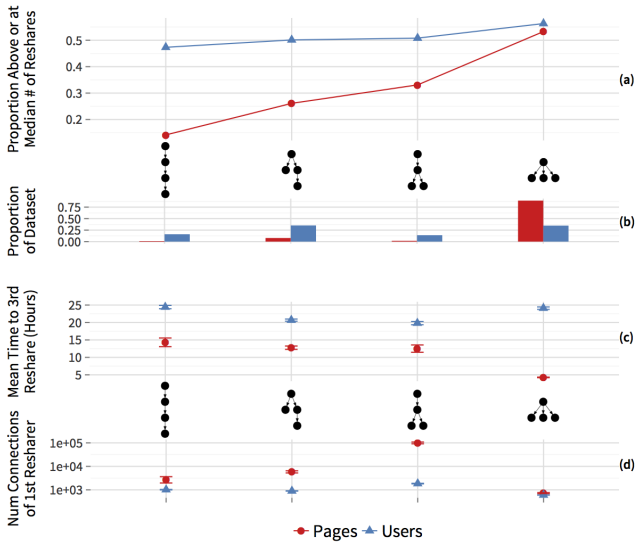


Figure 9: Shallow initial cascade structures are indicative of larger cascades. In contrast to page-started cascades, where the mean time to the 3rd reshare decreases with decreasing depth of the initial cascade, shallow cascades take a much longer time to form for user-started cascades. For these, the connections of the 1st resharer also significantly impacts the time to the 3rd resharer, especially when it receives two reshares before the original receives a second.

initiated by a popular page and achieving a large number of reshares directly from its fans. Interestingly, the configuration having the second and third reshares stemming from the first reshare correspond to having a first resharer with many connections, and indicating that the initial poster is less popular, be it a page or user (Figure 9d).

Curiously, for user-started cascades, the star configuration tends to grow into the largest cascades, but is also the slowest. It also tends to correspond to the first resharer having a low degree, both for page and user roots. One might speculate that this pattern is indicative of the item's appeal to less well-connected users, who also happen to be more likely to reshare. In fact, a median resharer has 35 fewer friends than someone who is active on the site nearly every day. Thus, an item's appeal, rather than the initial network structure, may drive the eventual cascade size in the long run.

4.3 Predicting cascade structure

The observations above naturally lead to the question of whether it is possible to predict future cascade structure. In particular, we aim to distinguish cascades that spread like a virus in a shallow forest fire-like pattern (Figure 3a) and cascades which spread in long, narrow string-like pattern (Figure 3c). As discussed earlier, this difference is related to the structural virality of a cascade and is quantified by the Wiener index. Here, we observe $k = 5$ reshares of a cascade and aim to predict whether the final cascade will have a Wiener index above or below the median. We obtain accuracy of 0.725 (F1 = 0.715, AUC = 0.796), while random guessing would, by construction, achieve accuracy of 0.5.

Temporal and structural features are most predictive of structure. For this task we expect structural features to be most important, while we expect temporal features not to be indicative of the cascade structure. However, when we train the model on individual classes of features we surprisingly find that both temporal and structural features are almost equally useful in predict-

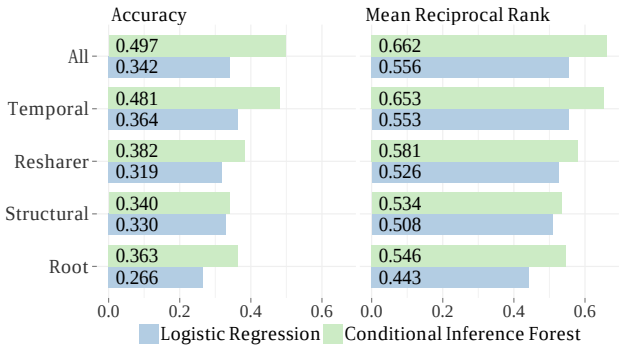


Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.

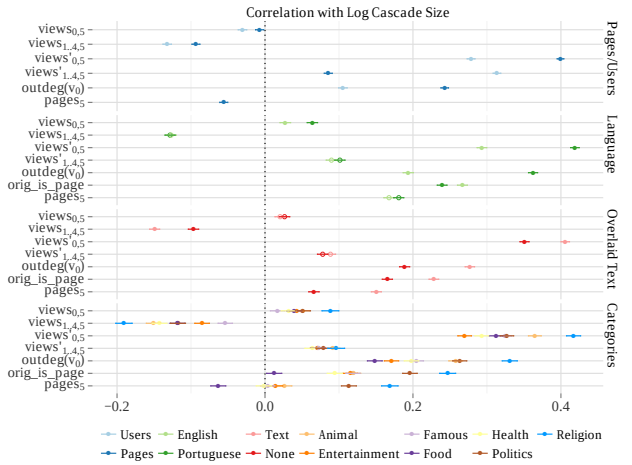


Figure 11: The initial exposure of the uploaded photo and initial reshares serve to differentiate datasets from one another, as can be seen by comparing the correlation coefficients of each feature with the log cascade size. Solid circles indicate significance at $p < 10^{-3}$, and lines through each circle indicate the 95% confidence interval.

ing cascade structure: 0.622 vs. 0.620. Nevertheless, structural features remain individually more accurate (≈ 0.58) and highly correlated ($0.161 \leq |r| \leq 0.255$) with the Wiener index. Individually, one temporal feature, $views'_{1..k-1,k}$, is slightly more accurate (0.602) compared to the best-performing structural feature, $outdeg(\hat{v}_0)$ (0.600), but is significantly less correlated (0.041 vs. -0.255). The two classes of features nicely complement each other, since when combined, accuracy increases to 0.72.

Cascade structure also becomes more predictable with increasing k. Like for cascade growth prediction, our prediction performance improves as we observe more of the cascade, with accuracy linearly increasing from 0.724 when k is 5 to 0.808 when k is 100. A linear relation also exists in the alternate task where we set the minimum cascade size R to be 100, varying k between 5 and 100.

Changes in feature importance. As we increase k, we find that the structural features become highly correlated with the Wiener index, suggesting that the initial shape of a cascade is a good indicator of its final structure. Rapidly growing cascades also result in final structures that are shallower—temporal features become more strongly correlated with the Wiener index as k increases. Unlike with cascade size, views were generally weakly correlated with structure, while content features had a weak, near-constant effect. Nonetheless, some of these features still provided reasonable performance in the prediction task.

User vs. page-started cascades. In predicting the shape of a cascade, we find that our overall prediction accuracy for pages is slightly higher (0.724) than for users (0.700). While using only structural features alone results in a higher prediction accuracy for users (0.643) than for pages (0.601), user and content features are significantly more predictive of cascade structure in the case of pages.

To sum up, we find that predicting the shape of a cascade is not as hard as one might fear. Nevertheless, predicting cascade size is still much easier than predicting cascade shape, though classifiers for either achieve non-trivial performance.

5. PREDICTABILITY & CONTENT

5.1 Controlling for cascade content

In our analyses thus far, we examined cascades of uploads of different photos, and tried to account for content differences by including photo and caption features. However, temporal and structural features may still capture some of the difference in content. Thus, we now study how well we can predict cascade size if we control for the content of the photo itself. We consider *identical photos* uploaded to Facebook by different users and pages, which is not a rare occurrence. We used an image matching algorithm to identify copies of the same image and place their corresponding cascades into clusters (983 clusters, $N_c = 38{,}073$, $N_r = 12{,}755{,}621$). As one might expect, even the same photo uploaded at different times by different users can fare dramatically differently; a cluster typically consists of a few or even a single cascade with a large number of reshares, and many smaller cascades with few reshares. The average Gini coefficient, a measure of inequality, is 0.787 ($\sigma = 0.104$) within clusters. Thus, a natural task is to try to predict the largest cascade within a cluster. For every cluster we select 10 random cascades, placing the accuracy of random guessing at 10%.

As shown in Figure 10, in all cases we significantly outperform the baseline. Using a random forest model, we can identify the most popular cascade nearly half the time (accuracy 0.497); a mean reciprocal rank of 0.662 indicates that this cascade also appears in the top two predicted cascades almost all the time.

In terms of feature importance we notice that best results are obtained using temporal features, followed by resharer, root node, and structural features. Essentially, if one upload of the photo is initially spreading more rapidly than other uploads of the same photo, that cascade is also likely to grow to be the largest. This points to the importance of landing in the right part of the network at the right time, as the same photo tends to have widely and predictably varying outcomes when uploaded multiple times.

5.2 Feature importance in context

Some features may be more or less important for our prediction tasks in different contexts. Figure 11 shows how several features correlate with log-transformed cascade size when conditioned on one of four different variables, including (1) source node type—user vs page, (2) language—English versus Portuguese, the two most common languages of cascade root nodes in our dataset, (3) whether text is overlaid on a photo—a common feature of recent Internet memes, and (4) content category. We determine content category by matching entities in photo captions to Wikipedia articles, and in turn articles to seven higher-level categories: animal, entertainment, politics, religion, famous people (excluding religious and political figures), food, and health.

Figure 11 shows that the initial rate of exposure of the uploaded photo is generally more important for page cascades than for user cascades ($views'_{0,5}$). This is likely due to the higher variance in the distribution of the number of followers for a user versus a page. For page cascades in our sample, the median number of followers is 73,855 with a standard deviation of 675,203, while for users at the root of cascades the median number of friends and subscribers is 1,042 with a standard deviation of 26,482. Though rate of exposure to the original photo is more important for pages, we see that rate of exposure to the initial reshares ($views'_{1..4,5}$) is much more important for user cascades.

The number and rate of views also act to differentiate topical categories, with religion having the highest correlation between views and cascade size. Correlation for the rate of views of the uploaded photo is also higher for those with a Portuguese-speaking root node as opposed to an English one. The feature $outdeg(v_0)$ indicates the ability of the root to broadcast content, and we see this playing an important role for page cascades, Portuguese content, photos with text, and religious photos. This indicates that much of the success of these cascades is related to the root nodes being directly connected to large audiences.

In addition to the analysis of Figure 11, we also examined how the features correlate with the structural virality of the final cascades. (Each of the reported correlation coefficient comparisons that follow are significant at $p < 10^{-3}$ using a Fisher transformation.) Photos relating to food differ significantly from all other categories in that features of the root, such as $outdeg(v_0)$, are less negatively correlated (>–0.18 vs. –0.11), and depth features, such as $depth_k^{avg}$, are less positively correlated (>0.18 vs. 0.11). This relationship also holds for English compared to Portuguese photos. While users with many friends or followers are more likely to generate cascades of larger size and greater structural virality, pages with many fans create cascades of larger size, although not necessarily greater virality (0.05 vs. –0.01). However, if the initial structure of a cascade is already deep, the final structure of the cascade is likely to have greater structural virality for both user and page-started cascades (>0.16). A user-started cascade whose initial reshares are viewed more quickly is also more likely to become viral than that for a page-started cascade (0.23 vs. 0.06).

6. DISCUSSION & CONCLUSION

This paper examines the problem of predicting the growth of cascades over social networks. Although predictive tasks of similar spirit have been considered in the past, we contribute a novel formulation of the problem which does not suffer from skew biases. Our formulation allows us to study predictability throughout the life of a cascade. We examine not only how the predictability changes as more and more of the cascade is observed (it improves), but also how predictable large cascades are if we only observe them initially (larger cascades are more difficult to predict). While some features, e.g., the average connection count of the first k resharers, have increasing predictive ability with increasing k, others weaken in importance, e.g., the connectivity of the root node. We find that the importance of features depends on properties of the original upload as well: the topics present in the caption, the language of the root node, as well as the content of the photo.

Despite the rich set of results we were able to obtain, there are some limitations to this study. Most importantly, the study was conducted entirely with Facebook data and only with photos. Still, one advantage of this is the scale of the medium; hundreds of millions of photos are uploaded to Facebook every day, and photos, more than other content types, tend to dominate reshares. This also gives us high-fidelity traces of how the photo moves within Facebook's ecosystem, which allows us to precisely overlay the spreading cascade over the social network. Moreover, we are able to identify uploads of the same photo and track them individually. This eliminates the concern of shares being driven by an external entity and only appearing to be spreading over the network. Instead, external drivers benefit our study by creating independent 'experiments' where the same photo gets multiple chances to spread, helping us control for the role of content in some of our experiments. Another disadvantage of our setup is that diffusion within Facebook is driven by the mechanics of the site. The distinction between pages and users is specific to Facebook, as are the mechansisms by which users interact with content, e.g., liking and resharing. Despite these limitations, we believe the results give general insights which will be useful in other settings.

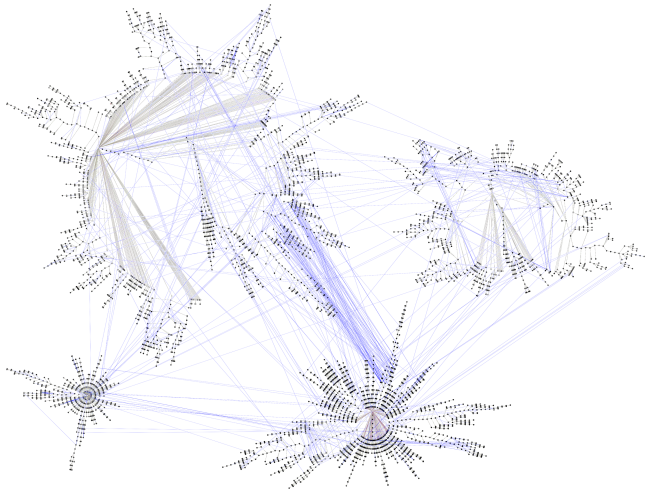


Figure 12: There is considerable overlap in friendship edges (blue) between four independent cascades of the same photo.

The present work only examines each cascade independently from others. Future work should examine interactions between cascades, both between different content competing for the same attention, and between the same content surfacing at different times and in different parts of the network. We found that when the same photo is uploaded at least 10 times, the largest cascade was twice as likely to be among the first 20% of uploads than the last 20%. Similarly, for photos uploaded 20 times, the largest cascade was 2.3 times as likely to be among the first 20% than the last. Figure 12 shows the friendship edges between users participating in different cascades of a single, specific photo. The high connectivity *between* different cascades demonstrates that users are likely being exposed to the same photo via different cascades, which could be a contributing factor in why earlier uploads of the same photo tend to generate larger cascade than later ones. Between-cascade dynamics like this should provide ample opportunities for further research.

Addressing questions like these will lead to a richer understanding of how information spreads online and pave the way towards better management of socially shared content and applications that can identify trending content in its early stages.

7. REFERENCES

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.

[2] A. Anderson, S. Goel, J. Hofman, and D. Watts. The structural virality of online diffusion. *Under review*.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[4] L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proc. WSDM*, 2013.

[5] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proc. WSDM*, 2011.

[6] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *Proc. EC*, 2009.

[7] J. Berger and K. L. Milkman. What makes online content viral. *J. Marketing Research*, 49(2):192–205, 2012.

[8] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *Proc. ICWSM*, 2010.

[9] P. A. Dow, L. A. Adamic, and A. Friggeri. The anatomy of large facebook cascades. In *Proc. ICWSM*, 2013.

[10] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers-predicting information cascades in microblogs. In *Proc. OSM*, 2010.

[11] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proc. EC*, 2012.

[12] B. Golub and M. O. Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proc. Natl. Acad. Sci.*, 2010.

[13] D. Gruhl, R. V. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. WWW*, 2004.

[14] M. Guerini, J. Staiano, and D. Albanese. Exploring image virality in google plus. *Proc. SocialCom*, 2013.

[15] T.-A. Hoang and E.-P. Lim. Virality and susceptibility in information diffusions. In *Proc. ICWSM*, 2012.

[16] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proc. WWW Companion*, 2011.

[17] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In *Proc. WWW Companion*, 2013.

[18] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *Proc. KDD*, 2010.

[19] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *Proc. CIKM*, 2012.

[20] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 2007.

[21] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proc. ICDM*, 2007.

[22] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci.*, 2008.

[23] Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 2013.

[24] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Proc. KDD*, 2012.

[25] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: LIWC 2001. 2001.

[26] S. Petrovic, M. Osborne, and V. Lavrenko. RT to win! predicting message propagation in twitter. In *Proc. ICWSM*, 2011.

[27] D. M. Romero, C. Tan, and J. Ugander. On the interplay between social and topical structure. In *Proceedings of the Seventh International Conference on Weblogs and Social Media (ICWSM)*, 2013.

[28] M. Salganik, P. Dodds, and D. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 2006.

[29] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 2010.

[30] O. Tsur and A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proc. WSDM*, 2012.

[31] D. J. Watts. *Everything is Obvious: How Common Sense Fails Us*. Crown, 2012.

[32] L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Sci. Rep.*, 2013.

[33] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *Proc. ICWSM*, 2010.

[34] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Proc. ICDM*, 2010.

Structural diversity in social contagion

Johan Ugander[a], Lars Backstrom[b], Cameron Marlow[b], and Jon Kleinberg[c,1]

[a]Center for Applied Mathematics and [c]Department of Computer Science, Cornell University, Ithaca, NY 14853; and [b]Facebook, Menlo Park, CA 94025

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved February 21, 2012 (received for review October 6, 2011)

The concept of contagion has steadily expanded from its original grounding in epidemic disease to describe a vast array of processes that spread across networks, notably social phenomena such as fads, political opinions, the adoption of new technologies, and financial decisions. Traditional models of social contagion have been based on physical analogies with biological contagion, in which the probability that an individual is affected by the contagion grows monotonically with the size of his or her "contact neighborhood"—the number of affected individuals with whom he or she is in contact. Whereas this contact neighborhood hypothesis has formed the underpinning of essentially all current models, it has been challenging to evaluate it due to the difficulty in obtaining detailed data on individual network neighborhoods during the course of a large-scale contagion process. Here we study this question by analyzing the growth of Facebook, a rare example of a social process with genuinely global adoption. We find that the probability of contagion is tightly controlled by the number of connected components in an individual's contact neighborhood, rather than by the actual size of the neighborhood. Surprisingly, once this "structural diversity" is controlled for, the size of the contact neighborhood is in fact generally a negative predictor of contagion. More broadly, our analysis shows how data at the size and resolution of the Facebook network make possible the identification of subtle structural signals that go undetected at smaller scales yet hold pivotal predictive roles for the outcomes of social processes.

social networks | systems

Social networks play host to a wide range of important social and nonsocial contagion processes (1–8). The microfoundations of social contagion can, however, be significantly more complex, as social decisions can depend much more subtly on social network structure (9–17). In this study we show how the details of the network neighborhood structure can play a significant role in empirically predicting the decisions of individuals.

We perform our analysis on two social contagion processes that take place on the social networking site Facebook: the process whereby users join the site in response to an invitation e-mail from an existing Facebook user (henceforth termed "recruitment") and the process whereby users eventually become engaged users after joining (henceforth termed "engagement"). Although the two processes we study formally pertain to Facebook, their details differ considerably; the consistency of our results across these differing processes, as well as across different national populations (*Materials and Methods*), suggests that the phenomena we observe are not specific to any one modality or locale.

The social network neighborhoods of individuals commonly consist of several significant and well-separated clusters, reflecting distinct social contexts within an individual's life or life history (18–20). We find that this multiplicity of social contexts, which we term structural diversity, plays a key role in predicting the decisions of individuals that underlie the social contagion processes we study.

We develop means of quantifying such structural diversity for network neighborhoods, broadly applicable at many different scales. The recruitment process we study primarily features small neighborhoods, but the on-site neighborhoods that we study in the context of engagement can be considerably larger. For small neighborhoods, structural diversity is succinctly measured by the number of connected components of the neighborhood. For larger neighborhoods, however, merely counting connected components fails to distinguish how substantial the components are in their size and connectivity. To determine whether the structural diversity of on-site neighborhoods is a strong predictor of on-site engagement, we evaluate several variations of the connected component concept that identify and enumerate substantial structural contexts within large neighborhood graphs. We find that all of the different structural diversity measures we consider robustly predict engagement. For both recruitment and engagement, structural diversity emerges as an important predictor for the study of social contagion processes.

Results

User Recruitment. To study the spread of Facebook as it recruits new members, we require information not just about Facebook's users but also about individuals who are not yet users. Thus, suppose that an individual A is not a user of Facebook; it is still possible to identify a set of Facebook users that A may know because these users have all imported A's e-mail address into Facebook. We define this set of Facebook users possessing A's e-mail address to be A's contact neighborhood in Facebook. This contact neighborhood is the subset of potential future friendship ties that can be determined from the presence of A's e-mail address (Fig. 1*A*). Whereas A may in fact know many other people on Facebook as well, such additional friendship ties remain unknown for individuals who do not choose to register and so cannot be studied as a predictor of recruitment. The e-mail contact neighborhoods we study are generally quite small, typically on the order of five or fewer nodes.

We can now study an individual's decision to join Facebook as follows. Facebook provides a tool through which its users can e-mail friends not on Facebook to invite them to join; such an e-mail invitation contains not only a presentation of Facebook and a profile of the inviter, but also a list of the other members of the individual's contact neighborhood. We analyze a corpus of 54 million such invitation e-mails, and the fundamental question we consider is the following: How does an individual's probability of accepting an invitation depend on the structure of his or her contact neighborhood?

Traditional hypotheses suggest that this probability should grow monotonically in the size of the contact neighborhood (3, 9, 10). What we find instead, however, is a striking stratification of acceptance probabilities by the number of connected components in the contact neighborhood (Fig. 1 *B–D* and Fig. S1). When going beyond component count, one may suspect that edge density has a significant impact on the recruitment conversion rate: Among the single-component neighborhoods of a given size, there is a considerable structural difference between neighborhoods connected as a tree and those connected as a clique. However, within the controlled conditional datasets of

Author contributions: J.U., L.B., C.M., and J.K. designed research; J.U., L.B., C.M., and J.K. performed research; J.U., L.B., C.M., and J.K. contributed new reagents/analytic tools; J.U., L.B., C.M., and J.K. analyzed data; and J.U. and J.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: kleinber@cs.cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1116502109/-/DCSupplemental.

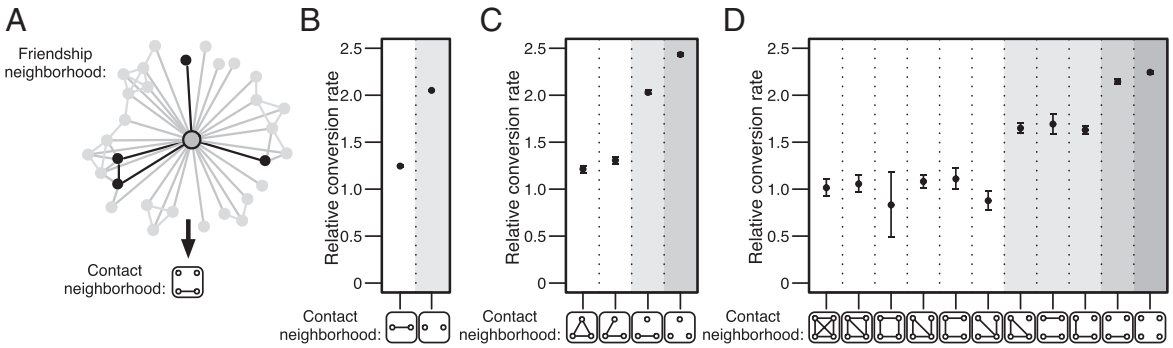


Fig. 1. Contact neighborhoods during recruitment. (*A*) An illustration of a small friendship neighborhood and a highlighted contact neighborhood consisting of four nodes and three components. (*B–D*) The relative conversion rates for two-node, three-node, and four-node contact neighborhood graphs. Shading indicates differences in component count. For five-node neighborhoods, see Fig. S1. Invitation conversion rates are reported on a relative scale, where 1.0 signifies the conversion rate of one-node neighborhoods. Error bars represent 95% confidence intervals and implicitly reveal the relative frequency of the different topologies.

one-component neighborhoods of sizes 4–6, we see that edge density has no discernible effect (Fig. 2*A*).

Moreover, we see that once component count is controlled for (Fig. 2*B*), neighborhood size is largely a negative indicator of conversion. In effect, it is not the number of people who have invited you, nor the number of links among them, but instead the number of connected components they form that captures your probability of accepting the invitation. Note that this analysis has been performed in aggregate and thus unavoidably reflects the decisions of different individuals. The ability to reliably estimate acceptance probabilities as a function of something as specific as the precise topology of the contact neighborhood is possible only because the scale of the dataset provides us with sufficiently many instances of each possible contact neighborhood topology (up through size 5).

We view the component count as a measure of "structural diversity," because each connected component of an individual's contact neighborhood hints at a potentially distinct social context in that individual's life. Under this view, it is the number of distinct social contexts represented on Facebook that predicts the probability of joining. We show that the effect of this structural diversity persists even when other factors are controlled for. In particular, the number of connected components in the contact neighborhood remains a predictor of invitation acceptance even when restricted to individuals whose neighborhoods are demographically homogeneous (in terms of sex, age, and nationality; Fig. S2), thus controlling for a type of demographic diversity that is potentially distinct from structural diversity. The component count also remains a predictor of acceptance even when we compare neighborhoods that exhibit precisely the same mixture of "bridging" and "embedded" links (Fig. S3), the key distinction in sociological arguments based on information novelty (19, 20).

For contact neighborhoods consisting of two nodes, we observe that the probability an invitation is accepted is much higher when the two nodes in the neighborhood are not connected by a link (hence forming two connected components, Fig. 1*B*) compared with when they are connected (forming one component). Is there a way to identify cases where people are likely to know each other, even if they are not linked on Facebook? The photo tagging feature on Facebook suggests such a mechanism. Photographs uploaded to Facebook are commonly annotated by users with "tags" denoting the people present in the photographs. We can use these tags to deduce whether two unlinked nodes in a contact neighborhood have been jointly tagged in any photos, a property we refer to as "co-tagging," which serves as an indication of a social tie through copresence at an event (21).

Using photo co-tagging, we find strong effects even in cases where the presence of a friendship tie is only implicit. If a contact neighborhood consists of two unlinked nodes that have nevertheless been co-tagged in a photo, then the invitation acceptance probability drops to approximately what it is for a neighborhood of two linked nodes (Fig. 2*C*). In other words, being co-tagged in a photo indicates roughly the same lack of diversity as being connected by a friendship link. We interpret this result as further evidence that diverse endorsement is key to predicting recruitment. Meanwhile, when the two nodes are friends, co-tags offer a proxy for tie strength, and we see that if the two nodes have also been co-tagged, then the probability of an accepted invitation decreases further. From this we can interpret tie strength as an

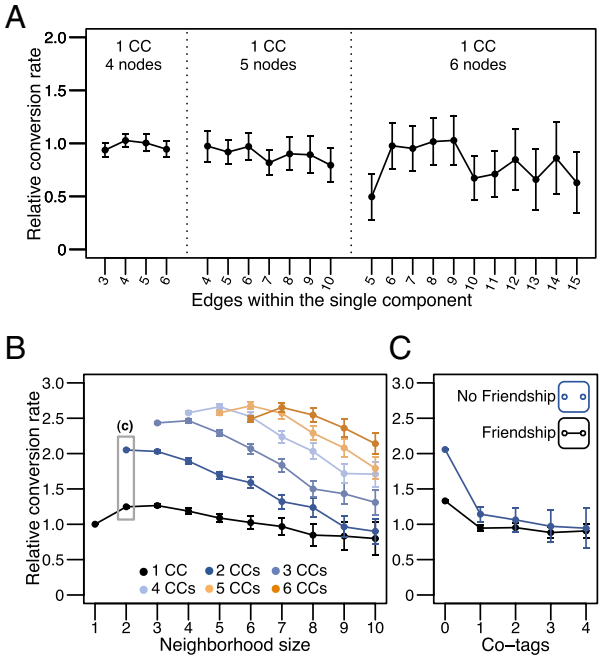


Fig. 2. Recruitment contact neighborhoods and component structure. (*A*) Conversion as a function of edge count neighborhoods with one connected component (1 CC) with four to six nodes, where variations in edge count predict no meaningful difference in conversion. (*B*) Conversion as a function of neighborhood size, separated by CC count. When component count is controlled for, size is a negative indicator of conversion. (*C*) Conversion as a function of tie strength in two-node neighborhoods, measured by photo co-tags, a negative indicator of predicted conversion. Recruitment conversion rates are reported on a relative scale, where 1.0 signifies the conversion rate of one-node neighborhoods. Error bars represent 95% confidence intervals.

extension of context, because two strongly tied nodes plausibly constitute an even less diverse endorsement neighborhood.

Finally, we study the position of the inviter within the neighborhood topologies. When studying recruitment, one might suspect that the structural position of the inviter—the person who extended the invitation—might signify differences in tie strength with the invitee and therefore might significantly affect the predicted conversion rate. We find that inviter position figures only slightly in the conversion rate (Fig. 3), with invitations stemming from a high-degree position in the contact neighborhood predicting only a slightly higher conversion rate than if the inviter is a peripheral node.

User Engagement. Participation in a social system such as Facebook is built upon a spectrum of social decisions, beginning with the decision to join (recruitment) and continuing on to decisions about how to choose a level of engagement. We now show how structural diversity also plays an analogous role in this latter type of decision process, studying long-term user engagement in the Facebook service. Whereas recruitment is a function of the complex interplay between multiple acts of endorsement, engagement is a function of the social utility a user derives from the service. Our study of engagement focuses on users who registered for Facebook during 2010, analyzing the diversity of their social neighborhoods 1 week after registration as a basis for predicting whether they will become highly engaged users 3 months later.

Users are considered engaged at a given time point if they have interacted with the service during at least 6 of the last 7 days. Facebook had 845 million monthly active users on December 31, 2011, and during the month of December 2011, an average of 360 million users were active on at least 6 out of the last 7 days. We define engagement on a weekly timescale to stabilize the considerable weekly variability of user visits. Our goal is therefore to predict whether a newly registered user will visit Facebook at least 6 of 7 days per week 3 months after registration.

Friendship neighborhoods on Facebook are significantly larger than the e-mail contact neighborhoods from our recruitment study. We focus our engagement study on a population of ∼10 million users who registered during 2010 and had assembled neighborhoods consisting of exactly 10, 20, 30, 40, or 50 friends 1 week after registration. For social network neighborhoods of this size, we find that a neighborhood containing a large number of connected components primarily indicates a large number of one-node components, or "singletons", and as such, it is not an accurate reflection of social context diversity.

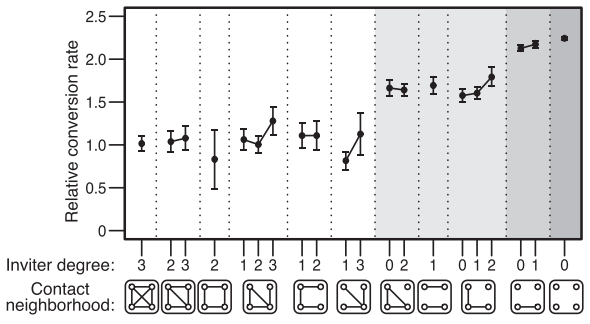


Fig. 3. Inviter position during recruitment. Shown is recruitment conversion as a function of neighborhood graph topology and inviter position in neighborhoods of size 4. The position of the inviter within the neighborhood graph is described exactly (up to symmetries) by node degree. Shading indicates differences in component count. Recruitment conversion rates are reported on a relative scale, where 1.0 signifies the conversion rate of one-node neighborhoods. Error bars represent 95% confidence intervals.

To address this, we evaluate three distinct parametric generalizations of component count. First, we measure diversity simply by considering only components over a certain size k. Second, we measure diversity by the component count of the *k-core* of the neighborhood graph (22), the subgraph formed by repeatedly deleting all vertices of degree less than k. Third, we define a measure that isolates dense social contexts by removing edges according to their *embeddedness*, the number of common neighbors shared by their two endpoints; intuitively this is an analog, for edges, of the type of node removal that defines the k-core. Adapting earlier work on embeddedness by Cohen (23), we define the *k-brace* of a graph to be the subgraph formed by repeatedly deleting all edges of embeddedness less than k and then deleting all single-node connected components. (Cohen's work was concerned with a definition equivalent to the largest connected component of the k-brace; because we deal with the full subgraph of all nontrivial components, it is useful to adapt the definitions as needed.) Examples of these three measures applied to a neighborhood graph are shown in Fig. 4 *A* and *B*, illustrating the

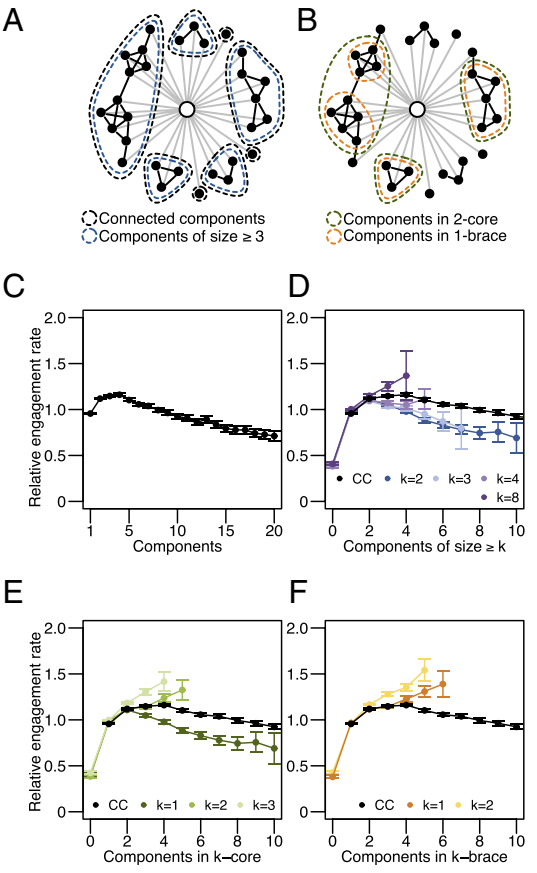


Fig. 4. Engagement and structural diversity for 50-node friendship neighborhoods. (*A*) Illustration of the connected components in a friendship neighborhood, delineating connected components and components of size ≥3. (*B*) Illustration of the k-core and the k-brace, delineating the connected components of the 2-core and the 1-brace. (*C*) Engagement as a function of connected component count. (*D*) Engagement as a function of the number of components of size ≥ k, for $k = 2, 3, 4, 8$, with connected component (CC) count shown for comparison. (*E*) Engagement as a function of k-core component count for $k = 1, 2, 3$, with CC count shown for comparison. (*F*) Engagement as a function of k-brace component count for $k = 1, 2$, with CC count shown for comparison. Engagement rates are reported on a relative scale, where 1.0 signifies the average conversion rate of all 50-node neighborhoods. All error bars are 95% confidence intervals. For other neighborhood sizes, see Fig. S4.

connected components of size 3 or greater, the connected components of the 2-core, and the connected components of the 1-brace. We see that the three parametric measures we evaluate differ measurably in how they isolate "substantial" social contexts.

The k-core component count for $k = 0$ is simply the component count of the original graph, the same as we analyzed when examining recruitment. For $k = 1$, the k-core component count is the count of nonsingleton components, whereas for $k = 2$, all tree-like components are discarded and the remaining components are counted. When considering the k-brace, observe that for all graphs the k-brace is a subgraph of the $(k + 1)$-core: indeed, because each node in the k-brace is incident to at least one edge, and each edge in the k-brace has embeddedness at least k, all nodes in the k-brace must have degree at least $k + 1$. It is therefore reasonable to compare the 1-brace to the 2-core. Both of these restrictions discard tree-like components, but the 1-brace will tend to break up components further than the 2-core does—the operation defining the 1-brace continues to cleave components in cases where sets of nodes forming triangles are linked together by unembedded edges or where a component contains cycles but no triangles. The notion of the k-core has been applied both to the study of critical phenomena in random graphs (24, 25) and to models of the Internet (26, 27), but to our knowledge the k-brace has not been studied extensively (see *SI Text* for some basic results on the k-brace and ref. 23 for analysis of a related definition).

When studying the structural diversity of 1-week Facebook friendship neighborhoods as a predictor of long-term engagement, simply counting connected components leads to a muddled view of predicted engagement (Fig. 4*C*). However, extending the notion of diversity according to any of the definitions above suffices to provide positive predictors of future long-term engagement. Specifically, when considering the components of the 1-brace, which removes small components and severs unembedded edges, we see that diversity (captured by the presence of multiple components) emerges as a significant positive predictor of future long-term engagement (Fig. 4*F*). We also see that the closely related 2-core component count is a clean predictor (Fig. 4*E*). Finally, if we consider simply the number of components of size k or larger in the original neighborhood (without applying the core or brace definitions), we see that small values of k are not enough (Fig. 4*D*); but even here, when k is increased to make the selection over components sufficiently astringent (in particular, when we count only components of size 8 or larger), a clean indicator of engagement again emerges.

When considering the k-brace, it is sufficient to consider the component count of the 1-brace for our purposes, but larger values of k may be useful for analyzing larger neighborhoods in other domains. We note that the presence of several components in the k-core and the k-brace is fundamentally limited by the size of the core/brace, and we perform a control of this potentially confounding factor (Fig. S5). The conventional wisdom for social systems such as Facebook is that their utility depends crucially upon the presence of a strong social context. Our findings validate this view, observing that the predicted engagement for users who lack any strong context (e.g., those who have zero components in their neighborhood 1-brace) is much lower than for those with such a context. Our analysis importantly extends this view, finding that the presence of multiple contexts introduces a sizable additional increase in predicted engagement.

A cruder approach to diversity might consider measuring diversity through the edge density of a neighborhood, figuring that sparse neighborhoods would be more varied in context. In Fig. 5 we see how this approach results in a complicated view where the optimal edge density for predicting engagement lies at an internal and size-dependent optimum. Given what our component analysis reveals, we interpret this observation as a superposition of two effects: Too few edges imply a lack of context (4) but too many edges imply a lacking diversity of contexts, with a nontrivial interior clearly dominating the boundary conditions. From Fig. 5 it also becomes clear that internal neighborhood structure is at least as important as size, with a 20-node neighborhood featuring a well-balanced density predicting higher conversion than a sparse or dense 50-node neighborhood.

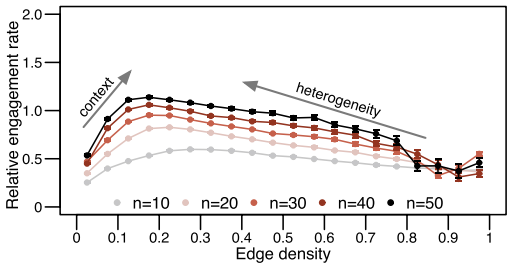


Fig. 5. Engagement as a function of edge density. For five different neighborhood sizes, $n = 10, 20, 30, 40, 50$, we see that when component count is not accounted for, an internal engagement optimum is observed, showing the combined forces of focused context and structural heterogeneity. Engagement rates are reported on a relative scale, where 1.0 signifies the average conversion rate of all 50-node neighborhoods. All error bars are 95% confidence intervals.

Discussion

Detailed traces of Facebook adoption provide natural sources of data for studying social contagion processes. Our analysis provides a high-resolution view of a massive social contagion process as it unfolded over time and suggests a rethinking of the underlying mechanics by which such processes operate. Rather than treating a person's number of neighbors as the crucial parameter, consider instead the number of distinct social contexts that these neighbors represent as the driving mechanism of social contagion.

The role of neighborhood diversity in contagion processes suggests interesting further directions to pursue, both for mathematical modeling and for potential broader applications. Mathematical models in areas including interacting particle systems (28, 29) and threshold contagion (3, 30) have explored some of the global phenomena that arise from contagion processes in networks for which the behavior at a given node has a nontrivial dependence on the full set of behaviors at neighboring nodes. Neighborhood diversity could be naturally incorporated into such models by basing the underlying contagion probability, for example, on the number of connected components formed by a node's affected neighbors. It then becomes a basic question to understand how the global properties of these processes change when such factors are incorporated.

More broadly, across a range of further domains, these findings suggest an alternate perspective for recruitment to political causes, the promotion of health practices, and marketing; to convince individuals to change their behavior, it may be less important that they receive many endorsements than that they receive the message from multiple directions. In this way, our findings propose a potential revision of core theories for the roles that networks play across social and economic domains.

Materials and Methods

Recruitment Data Collection. Here we discuss details of the e-mail recruitment data. All user data were analyzed in an anonymous, aggregated form. The contact neighborhood individuals included in invitation e-mails are limited to nine in number, and so we have restricted our analysis to neighborhoods (inviter plus contact importers) of 10 nodes or less. In cases with more than nine candidate "other people you may know," the invitation tool selects a randomized subset of nine for inclusion in the e-mail.

We conditioned our data collection upon several criteria. First, we considered only first invitations to join the site. Subsequent invitations to an e-mail address are handled differently by the invitation tool, and so we have not included them in our study. Second, we considered only invitations where

the inviter invited at most 20 e-mail addresses on the date of the invitation. This conditioning is meant to omit invitation batches where the inviter opted to "select all" within the contact import tool and focuses our investigation on socially selective invitations.

Invitations were sent during an 11-week period spanning July 12, 2010 to September 26, 2010. An e-mail address was considered to have converted to a registered user account if the address was registered for an account within 14 days of the invitation, counting both individuals who signed up via links provided in the invitation e-mail and users who signed up by visiting the Facebook website directly within 14 days. Only contact import events that occurred before the invitation event are considered. Likewise, only friendship edges that existed before the invitation event are considered to be part of the neighborhood.

Many of the findings we investigate are governed by complex nonlinear effects, which make traditional regression controls generally inadequate. In an attempt to control for confounding signals in our data, several parallel observation groups were maintained, against which all findings were validated. As a means of capturing potential artifacts from duplicitous private/business e-mail address use, a first such validation group was constructed by conditioning upon e-mail invitations sent to a small set of common and commonly private e-mail providers: Hotmail, Yahoo!, Gmail, AOL, and Yahoo! France. As a means of observing any differences between already established and growing Facebook markets, two parallel validation groups were constructed to observe established markets (United States) and emerging Facebook markets (Brazil, Germany, Japan, and Russia), classified by the most recently resolved country of login for the inviting Facebook account. Whereas invitation conversion rates were generally higher in emerging markets, none of the conditional datasets were observed to deviate from the complete dataset with regard to internal structural findings.

Highly sparse neighborhoods were a very common occurrence in these data, owing to the fact that the neighborhoods we study here are only partial observations of an individual's actual connection to Facebook. We are able to infer links only to those site users who have used the contact importer tool and maintain active e-mail communication with the e-mail address in question, criteria that induce a sampled subgraph that we then observe. The probability of sampling an edge uniformly at random in any neighborhood with low edge density is therefore quite low, and the probability that all sampled nodes come from the same cluster within a clustered neighborhood is lower still. From the perspective of communication multiplexity (31), we should in fact expect that our randomly induced subgraph sample is biased toward strongly connected ties that tend to communicate on multiple mediums, but this expectation is not at issue with our results. The real matter of the fact is that contact neighborhoods where the induced subgraph consists of a single connected component are likely to come from *very* tightly connected neighborhood graphs.

Although the contact importer tool and invitation tool are prominently featured as part of the new user experience on Facebook, they are also heavily used by experienced users of the site: The median site age of an inviter in our dataset was 262 days. Although e-mail invitations constitute only a small portion of Facebook's growth, they provide a valuable window into the otherwise invisible growth process of the Facebook product.

For the analysis of photo co-tags, only co-tags since January 1, 2010 were considered.

Engagement Data Collection. We consider users *engaged* at a given time point if they have interacted with the application during at least 6 of the last 7 days. As with any measure of user behavior, this metric is a heuristic merely meant to approximate a broader notion of involvement on the site. Highly engaged users who do not access the Internet on weekends will never qualify as "six-plus engaged," whereas users who simply log in on a daily basis to check their messages will qualify. Our analysis is restricted to the population level, so such confounders are not a problem.

Due to the technical nature of how engagement data are stored at Facebook, it is impractical to retrieve six-plus engagement measures for dates exactly 3 months after registration. As an appropriate surrogate, we consider the six-plus engagement of users on the first day of their third calendar month as users.

ACKNOWLEDGMENTS. We thank M. Macy, J. Fowler, D. Watts, and S. Strogatz for comments. This research has been supported in part by a MacArthur Foundation Fellowship and National Science Foundation Grants IIS-0705774, IIS-0910664, CCF-0910940, and IIS-1016099.

1. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Phys Rev Lett* 86:3200–3203.
2. Newman ME, Watts DJ, Strogatz SH (2002) Random graph models of social networks. *Proc Natl Acad Sci USA* 99(Suppl 1):2566–2572.
3. Dodds PS, Watts DJ (2004) Universal behavior in a generalized model of contagion. *Phys Rev Lett* 92:218701.
4. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: Membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D (Association for Computing Machinery, New York), pp 44–54.
5. Kearns M, Suri S, Montfort N (2006) An experimental study of the coloring problem on human subject networks. *Science* 313:824–827.
6. Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. *J Consum Res* 34:441–458.
7. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *N Engl J Med* 357:370–379.
8. Sun E, Rosenn I, Marlow C, Lento T (2009) Gesundheit! Modeling contagion through Facebook news feed. *Proceedings of the AAAI International Conference on Weblogs and Social Media*, eds Adar E, et al. (Association for the Advancement of Artificial Intelligence, Menlo Park, CA), pp 146–153.
9. Schelling T (1971) Dynamic models of segregation. *J Math Sociol* 1:143–186.
10. Granovetter M (1978) Threshold models of collective action. *Am J Sociol* 83: 1420–1443.
11. Burt R (1987) Social contagion and innovation: Cohesion versus structural equivalence. *Am J Sociol* 92:1287–1335.
12. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. *Science* 311:88–90.
13. Centola D, Eguíluz V, Macy M (2007) Cascade dynamics of complex propagation. *Physica A* 374:449–456.
14. Centola D, Macy M (2007) Complex contagions and the weakness of long ties. *Am J Sociol* 113:702–734.
15. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. *Nature* 446: 664–667.
16. Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc Natl Acad Sci USA* 106: 21544–21549.
17. Fowler JH, Christakis NA (2010) Cooperative behavior cascades in human social networks. *Proc Natl Acad Sci USA* 107:5334–5338.
18. Simmel G (1955) *Conflict and the Web of Group Affiliations*, eds trans Wolff K, Bendix R (Free Press, Glencoe, IL).
19. Granovetter M (1973) The strength of weak ties. *Am J Sociol* 78:1360–1380.
20. Burt R (1992) *Structural Holes: The SocialStructure of Competition* (Harvard Univ Press, Cambridge, MA).
21. Crandall D, et al. (2010) Inferring social ties from geographic coincidences. *Proc Natl Acad Sci USA* 107:22436–22441.
22. Bollobás B (2001) *Random Graphs* (Cambridge Univ Press, Cambridge, UK), 2nd Ed, p 150.
23. Cohen JD (2008) Trusses: Cohesive subgraphs for social network analysis. *National Security Agency Technical Report* (National Security Agency, Fort Meade, MD).
24. Luczak T (1991) Size and connectivity of the k-core of a random graph. *Discrete Math* 91:61–68.
25. Janson S, Luczak MJ (2007) A simple solution to the k-core problem. *Random Struct Algo* 30:50–62.
26. Alvarez-Hamelin JI, Dall'Astra L, Barrat A, Vespignani A (2006) Large scale networks fingerprinting and visualization using the k-core decomposition. *Adv Neural Inf Process Syst* 18:41–50.
27. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of Internet topology using k-shell decomposition. *Proc Natl Acad Sci USA* 104:11150–11154.
28. Liggett T (1985) *Interacting Particle Systems* (Springer, Berlin).
29. Durrett R (1995) *Ten Lectures on Particle Systems* (Springer, Berlin).
30. Mossel E, Roch S (2007) On the submodularity of influence in social networks. *Proceedings of the ACM Symposium on Theory of Computing*, eds Johnson DS, Feige U (Association for Computing Machinery, New York), pp 128–134.
31. Haythornthwaite C, Wellman B (1998) Work, friendship, and media use for information exchange in a networked organization. *J Am Soc Inf Sci* 49:1101–1114.

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Structural Virality of Online Diffusion

Sharad Goel, Ashton Anderson, Jake Hofman, Duncan J. Watts

To cite this article:
Sharad Goel, Ashton Anderson, Jake Hofman, Duncan J. Watts (2015) The Structural Virality of Online Diffusion. Management Science

Published online in Articles in Advance 22 Jul 2015

. http://dx.doi.org/10.1287/mnsc.2015.2158

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–17
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2015.2158

The Structural Virality of Online Diffusion

Sharad Goel, Ashton Anderson
Stanford University, Stanford, California, 94305 {scgoel@stanford.edu, ashton@cs.stanford.edu}

Jake Hofman, Duncan J. Watts
Microsoft Research, New York, New York 10016 {jmh@microsoft.com, duncan@microsoft.com}

Viral products and ideas are intuitively understood to grow through a person-to-person diffusion process analogous to the spread of an infectious disease; however, until recently it has been prohibitively difficult to directly observe purportedly viral events, and thus to rigorously quantify or characterize their structural properties. Here we propose a formal measure of what we label "structural virality" that interpolates between two conceptual extremes: content that gains its popularity through a single, large broadcast and that which grows through multiple generations with any one individual directly responsible for only a fraction of the total adoption. We use this notion of structural virality to analyze a unique data set of a billion diffusion events on Twitter, including the propagation of news stories, videos, images, and petitions. We find that across all domains and all sizes of events, online diffusion is characterized by surprising structural diversity; that is, popular events regularly grow via both broadcast and viral mechanisms, as well as essentially all conceivable combinations of the two. Nevertheless, we find that structural virality is typically low, and remains so independent of size, suggesting that popularity is largely driven by the size of the largest broadcast. Finally, we attempt to replicate these findings with a model of contagion characterized by a low infection rate spreading on a scale-free network. We find that although several of our empirical findings are consistent with such a model, it fails to replicate the observed diversity of structural virality, thereby suggesting new directions for future modeling efforts.

Keywords: Twitter; diffusion; viral media
History: Received August 14, 2013; accepted November 26, 2014, by Lorin Hitt, information systems.
Published online in *Articles in Advance*.

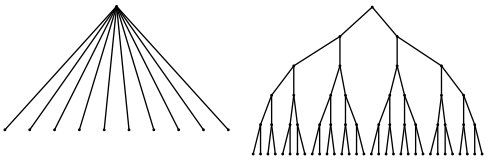
1. Introduction

When a piece of online media content—say, a video, an image, or a news article—is said to have "gone viral," it is generally understood not only to have rapidly become popular, but also to have attained its popularity through some process of person-to-person contagion, analogous to the spread of a biological virus (Anderson and May 1991). In many theoretical models of adoption (Coleman et al. 1957, Bass 1969, Mahajan and Peterson 1985, Valente 1995, Bass 2004, Toole et al. 2012), in fact, this analogy is made explicit: an "infectious agent"—whether an idea, a product, or a behavior—is assumed to spread from "infectives" (those who have it) to "susceptibles" (those who do not) via some contact process, where susceptibles can then be infected with some probability.[1] Both intuitively and also in formal theoretical models, therefore, the notion of viral spreading implies a rapid, large-scale increase in adoption that is driven largely, if not exclusively, by peer-to-peer spreading. Clearly, however, viral spreading is not the only mechanism by which a piece of content can spread to reach a large population. In particular, mass media or marketing efforts rely on what might be termed a "broadcast" mechanism, meaning simply that a large number of individuals can receive the information directly from the same source. As with viral events, broadcasts can be extremely large—the Superbowl attracts over 100 million viewers, while the front pages of the most popular news websites attract a similar number of daily visitors—and hence the mere observation that something is popular, or even that it became so rapidly, is not sufficient to establish that it spread in a manner that resembles social contagion.

Figure 1 schematically illustrates these two stylized modes of distribution—broadcast and viral—where the former is dominated by a large burst of adoptions from a single parent node, and the latter comprises a multigenerational branching process in which any one node directly "infects" only a few others. Although the stylized patterns in Figure 1 are intuitively plausible and also easily distinguishable from one another, differentiating systematically

[1] Even models of social contagion that do not correspond precisely to the mechanics of biological infectious disease (for example, "threshold models" (Granovetter 1978) make different assumptions regarding the nonindependence of sequential contacts with infectives (Lopez-Pintado and Watts 2008)) assume some form of person-to-person spread (Watts 2002, Kempe et al. 2003, Dodds and Watts 2004).

Figure 1 **A Schematic Depiction of Broadcast vs. Viral Diffusion, Where Nodes Represent Individual Adoptions and Edges Indicate Who Adopted from Whom**



between broadcast and viral diffusion requires one, in effect, to characterize the fine-grained structure of viral diffusion events. Yet, in spite of a large theoretical and empirical literature on the diffusion of information and products, relatively little is known about their structural properties, in part because the requisite data have not been available until very recently, and in part because the concept of virality itself has not been formulated previously in an explicitly structural manner. Classical diffusion studies (Coleman et al. 1957, Rogers 1962, Bass 1969, Valente 1995, Young 2009, Iyengar et al. 2010), for example, typically had access to only aggregate diffusion data, such as the cumulative number of adoptions of a product, technology, or idea over time (Fichman 1992). In such cases, the observation of an S-shaped adoption curve—indicating a period of rapid growth followed by saturation—is typically interpreted as evidence of social contagion (Rogers 1962); however, S-shaped adoption curves may also arise from broadcast distribution mechanisms such as marketing or mass media (Van den Bulte and Lilien 2001). Compounding the difficulty, real diffusion events are unlikely to conform precisely to either of these conceptual extremes. In a highly heterogeneous media environment (Walther et al. 2010, Wu et al. 2011), where any given piece of content can spread via email, blogs, and social networking sites as well as via more traditional offline media channels, one would expect that popular content might have benefited from some possibly complicated combination of broadcasts and interpersonal spreading.

To understand the underlying structure of an event, therefore, one must reconstruct the full adoption cascade, which in turn requires observing both individual-level adoption decisions and also the social ties over which these adoptions spread. Only recently have data satisfying these requirements become available, as a result of online behavior such as blogging (Adar and Adamic 2005, Yang and Leskovec 2010), e-commerce (Leskovec et al. 2006), multiplayer gaming (Bakshy et al. 2009), and social networking (Sun et al. 2009, Yang and Counts 2010, Bakshy et al. 2011, Petrovic et al. 2011, Goel et al. 2012, Hoang and Lim 2012, Tsur and Rappoport 2012, Kupavskii et al. 2012, Jenders et al. 2013, Ma et al. 2013).

A second empirical challenge in measuring the structure of diffusion events, which has in fact been highlighted by these recent studies, is that the vast majority of cascades—over 99%—are tiny and terminate within a single generation (Goel et al. 2012). Large and potentially viral cascades are therefore necessarily very rare events; hence, one must observe a correspondingly large number of events to find just one popular example, and many times that number to observe many such events. As we will describe later, in fact, even moderately popular events occur in our data at a rate of only about one in a thousand, whereas "viral hits" appear at a rate closer to one in a million. Consequently, to obtain a representative sample of a few hundred viral hits—arguably just large enough to estimate statistical patterns reliably—one requires an initial sample on the order of a billion events, an extraordinary data requirement that is difficult to satisfy even with contemporary data sources.

In this paper, we make three distinct but related contributions to the understanding of the structure of online diffusion events. First, we introduce a rigorous definition of *structural virality* that quantifies the intuitive distinction between broadcast and viral diffusion and allows for interpolation between them. As we explain in more detail below, our definition is couched exclusively in terms of observed patterns of adoptions, not on the details of the underlying generative process. Although this approach may seem counterintuitive in light of our opening motivation (which does make reference to generative models), the benefit is that the resulting measure does not depend on any modeling assumptions or unobserved properties, and hence can be applied easily in practice. Also importantly, by treating structural virality as a continuously varying quantity, we skirt any categorical distinctions between completely "broadcast" and "viral" events, allowing instead for open-ended and fine-grained distinctions between these two extremes; that is, events can be more or less structurally viral without imposing any particular threshold for becoming or "going" viral.

Our second contribution is to apply this measure of structural virality to investigate the diffusion of nearly a billion news stories, videos, pictures, and petitions on the microblogging service Twitter. To date, most studies directly documenting person-to-person diffusion have been limited to a small set of highly viral products (Liben-Nowell and Kleinberg 2008, Dow et al. 2013), leaving open the possibility that such hand-selected events are astronomically rare and not representative of viral diffusion more generally. In contrast, by systematically exploring the structural properties of a billion events on Twitter, we aim to estimate the frequency of structurally viral cascades, quantify the diversity in the structure of cascades,

and investigate the relationship between cascade size and structure. It could be, for example, that the most popular content is also extremely viral, but equally it could be that successful products are mostly driven by mass media (i.e., a single large broadcast) or by some combination of broadcasts and word of mouth. Depending on the relative importance of broadcast versus viral diffusion in driving popularity, that is, the relationship between popularity and structural virality could be positive (larger events are dominated by viral spreading), negative (larger events are dominated by broadcasts), or neither (all events regardless of size exhibit a similar mix of broadcasts and virality, which scale together). Applying our structural virality measure to a representative sample of successful cascades, we find evidence for the third possibility, namely, that the correlation between popularity and virality is generally low. Moreover, for any given size (equivalent popularity), structural virality is extremely diverse: cascades can range between pure "broadcasts," in the sense that all adopters receive the content from the same source, and highly "viral," in the sense of comprising multigenerational branching structures.

The third contribution of this paper is to compare our empirical observations of cascade structure to predictions from a series of simple generative models of diffusion. Specifically, we conduct large-scale simulations of a simple disease-like contagion model, similar to the original Bass (1969) model of product adoption, on a network comprising 25 million nodes. In the simplest variant, we assume that the infectiousness of the "disease" is a constant, and the network on which it spreads is an Erdős–Rényi (ER) random graph. In successively more complicated variants, we allow the infectiousness to vary, or the network to be "scale free" (i.e., where the number of neighbors can vary from tens to tens of millions), or both. Because large diffusion events are so rare, we also conduct on the order of 1 billion simulations per parameter setting, necessitating over 100 billion simulations in total. We find that although our simplest models are incapable of replicating even the most general features of our empirical data, a still-simple model comprising constant infectiousness and scale-free degree distribution can capture many, but not all, of the observed features. We conclude with some suggestions for future modeling efforts.

2. Defining Structural Virality

We now turn to our first goal of defining structural virality. Before proceeding, we reemphasize that our notion of structural virality is intended to complement, not substitute for, the many existing generative models of viral propagation and their associated parameters (Bass 1969, Granovetter 1978, Watts 2002, Kempe et al. 2003, Dodds and Watts 2004). To clarify, generative models attempt to describe the underlying diffusion mechanism itself—for example, as a function of the intrinsic infectiousness of the object that is spreading, or of the properties of the contact process or the network over which the diffusion occurs, or of the timescales associated with adoption. By contrast, our notion of structural virality is concerned exclusively with characterizing the structure of the observable adoption patterns that arise from some unobserved generative process. Naturally, the particular value of structural virality associated with some event will in general depend on the underlying generative process—as indeed we will demonstrate in §5, where we introduce and study several such models. Importantly, however, our desired *definition* of structural virality should not depend on these particulars. In other words, regardless of what contagion process is (assumed to be) responsible for some piece of content spreading or what network it is spreading over, the end result is some pattern of adoptions that exhibits some structure, and our goal is to characterize a particular property of that structure.

Recalling also that our goal is to disambiguate between the broadcast and multigenerational branching schematics depicted in Figure 1, we first lay out some intuitively reasonable criteria that we would like any such metric to exhibit. First, for a fixed total number of adoptions in a cascade, structural virality should increase with the branching factor of the structure: specifically, it should be minimized for the broadcast structure on the left of Figure 1 and should be relatively large for structures with a high branching factor, as on the right of Figure 1. Second, for a fixed branching factor, structural virality should increase with the number of generations (i.e., depth) of the cascade; that is, all else equal, larger branching structures should be more structurally viral than smaller ones. Finally, and in contrast with multigenerational branching structures, larger broadcasts should not be any more structurally viral than smaller broadcasts; hence we require that, for the extreme case of a pure broadcast, structural virality be approximately independent of size.

A natural choice for such a metric is simply the number of generations, or depth, of the cascade. Indeed, after size, depth is one of the most widely reported summary statistics of diffusion cascades (Liben-Nowell and Kleinberg 2008, Goel et al. 2012, Dow et al. 2013). One problem with depth, however, is that a single, long chain can dramatically affect the measure. For example, a large broadcast with just one, long, multigenerational branch has large depth, even though we would not intuitively consider it to be structurally viral. To correct for this issue, one could

instead consider the average depth of nodes (i.e., the average distance of nodes from the root). This average depth measure alleviates the problem of a handful of nonrepresentative nodes skewing the metric, and intuitively distinguishes between broadcasts and multigenerational chains. Even this measure, however, fails in certain cases. Notably, if an idea or product traverses a long path from the root and then is broadcast out to a large group of adopters, the corresponding cascade would have high average depth (since most adopters are far from the root) even though most adoptions in this case are the result of a single influential node.

Addressing the shortcomings of both depth and average depth, we focus our attention on a classical graph property studied originally in mathematical chemistry (Wiener 1947), where it is known as the "Wiener index." Specifically, we define structural virality $\nu(T)$ as the average distance between all pairs of nodes in a diffusion tree T; that is, for $n > 1$ nodes,

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}, \tag{1}$$

where d_{ij} denotes the length of the shortest path between nodes i and j.[2] Equivalently, $\nu(T)$ is the average depth of nodes, averaged over all nodes in turn acting as a root.

Our metric $\nu(T)$ provides a continuous measure of structural virality, with higher values indicating that adopters are, on average, farther apart in the cascade, and thus suggesting an intuitively viral diffusion event. In particular, as with depth and average depth, over the set of all trees on n nodes $\nu(T)$ is minimized on the star graph (i.e., the stylized broadcast model in Figure 1) where $\nu(T) \approx 2$. Moreover, a complete k-ary tree (as in Figure 1 with $k = 2$) has structural virality approximately proportional to its height; hence, structural virality will be maximized for structures that are large and that become that way through many small branching events over many generations.[3]

Although $\nu(T)$ satisfies some basic requirements of theoretical plausibility, as with the other candidate measures we discussed it is possible to construct hypothetical examples for which the corresponding numerical values are at odds with the motivating intuition. For example, a graph comprised of two stars connected by a single, long path has large $\nu(T)$ but would not intuitively be considered viral. Whether or not such pathological cases appear with any meaningful frequency is, however, largely an empirical matter, and hence the utility of the metric must ultimately be evaluated in the context of real examples, which we discuss in detail below as well as in Appendix B.

3. Data and Methods

Our primary analysis is based on approximately 1 billion diffusion events on Twitter, where an event constitutes the independent introduction of a piece of content into the social network—including videos, images, news stories, and petitions—along with all subsequent repostings of the same item.[4] Specifically, we include in our data all tweets posted on Twitter that contained URLs pointing to one of several popular websites over a 12 month period, from July 2011 to June 2012.[5] In total, we observe roughly 622 million unique pieces of content; however, because individual pieces of content can be posted by multiple users, we observe approximately 1.2 billion "adoptions" (i.e., posting of content). Although our data are not a total sample of Web content that is shared on Twitter,[6] they do include the vast majority and hence are essentially unbiased at least with respect to Tweets linking to Web content.[7] Importantly for our conclusions, our sample also exhibits considerable diversity both with respect to production and consumption. For example, a typical online video is likely to have been produced and distributed by

[2] Naive computation of $\nu(T)$ requires $O(n^2)$ time; however, as discussed in Appendix B, a more sophisticated approach yields a linear-time algorithm (Mohar and Pisanski 1988), facilitating computation on very large cascades.

[3] Somewhat more precisely, for any branching ratio $k << n$, $\nu(T)$ increases with size n, whereas for $k \approx n$ (i.e., pure broadcasts) it does not; hence, increasing popularity corresponds to increasing structural virality only when it arises from "viral" spreading, not merely from larger broadcasts.

[4] We use the term "reposting" rather than the more conventional "retweet" because individuals frequently repost content that they receive from another user without using the explicit retweet functionality provided by Twitter, or even acknowledging the source of the content.

[5] For news those websites include bbc.co.uk, cnn.com, forbes.com, nytimes.com, online.wsj.com, guardian.co.uk, huffingtonpost.com, news.yahoo.com, usatoday.com, telegraph.co.uk, and msnbc.msn.com. For video they include youtube.com, m.youtube.com, youtu.be, vimeo.com, livestream.com, twitcam.livestream.com, ustream.tv, twitvid.com, mtv.com, and vh1.com. For images they include twitpic.com, instagr.am, instagram.com, yfrog.com, p.twimg.com, twimg.com, i.imgur.com, imgur.com, img.ly, and flickr.com. For petitions they include change.org, twitition.com, kickstarter.com.

[6] URLs and redirects were dereferenced from original tweets, and extraneous query parameters were removed from URLs to identify multiple versions of identical content. To avoid left censoring of our data (i.e., missing the initial postings of a URL), we look for occurrences of the URLs during the month prior to our analysis period and only include in our sample instances where the first observation does not appear before July 1, 2011. To avoid right censoring, we restrict to tweets introduced prior to June 30, 2012, but continue tracing the diffusion of these tweets through July 31, 2012.

[7] It is of course possible that Tweets containing links to Web content are systematically different from other Tweets in ways that might affect our conclusions. For this reason, in Appendix D we conduct a separate analysis of tweets containing long hashtags, which are unlikely to diffuse outside of Twitter, finding qualitatively similar results.

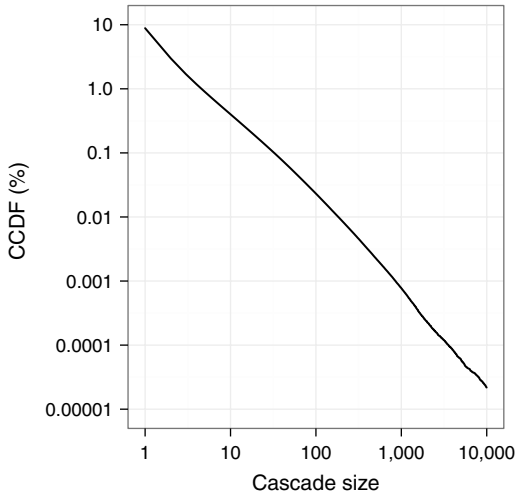
an amateur videographer uploading his or her own work onto YouTube, whereas an article appearing in a major news outlet was likely written by a professional reporter. Moreover, the experience of watching a video is quite distinct from that of reading a news article, both in terms of the time and effort required on the part of the consumer and also their goals—for example, to be entertained versus informed—in doing so. Due in part to these qualitative differences on both the supply and also demand sides of the market for media, we find large quantitative differences in the frequency of the four domains; specifically, images and videos are far more numerous than news stories, and petitions are by far the least numerous. For similar reasons, therefore, one might also expect qualitatively distinct sharing mechanisms to dominate in different domains, leading to different patterns both of popularity and also structural virality.

To evaluate the structure of online diffusion, for each independent introduction of a unique piece of content in our data we construct a corresponding diffusion "tree" that traces each adoption back to a single "root" node, namely, the user who introduced that particular piece of content.[8] Specifically, for each observation of a URL whose diffusion we seek to trace, we record (1) the adopter (i.e., the identity of the user who posted the content); (2) the adoption time (i.e., the time at which the content was posted); and (3) the identities of all users the adopter follows—hereafter referred to as the adopter's "friends"—from whom the adopter could conceivably have learned about the content. For each such event, we first determine whether at least one of the adopter's friends adopted the same piece of content previously. If no such friend exists, then the adopter is labeled a "root" of the resulting diffusion tree; otherwise, the friend who adopted the content most recently before the focal adopter—and who is most likely to have exposed the focal user to the content—is labeled the focal adopter's "parent." Although there is at times genuine ambiguity in determining the proximate cause of an adoption, in many cases adopters explicitly credit another individual in their tweet, allowing us to accurately infer an adopter's parent in approximately 95% of instances (see Appendix C for details of the tree construction algorithm and the associated evaluation procedure).

4. Results

Consistent with previous work (Bakshy et al. 2011, Goel et al. 2012), we find that the average size of these diffusion trees (also referred to interchangeably as "cascades" or "diffusion events") is 1.3—meaning that for every 10 introductions of content, there are on average three additional downstream adoptions. More strikingly, and as noted in Goel et al. (2012), we also find that the vast majority of cascades terminate within a single generation; specifically, about 99% of adoptions are accounted for either by the root nodes themselves or by the immediate followers of root nodes. As noted previously (Goel et al. 2012), however, the preponderance of small and shallow events does not rule out the possibility that large, structurally interesting events do occur, only that they occur sufficiently infrequently so as not to be observed even in relatively large data sets. Exploiting the fact that we have a much larger data set than in previous studies—over a billion observations in our initial sample—we therefore now focus exclusively on the subsample of rare events that qualify as large, and hence have the potential to be structurally interesting. Specifically, hereafter we restrict attention to the 0.025% of diffusion trees containing at least 100 nodes (Figure 2), a requirement that leaves us with roughly 1 out of every 4,000 cascades, and thus reduces the number of cascades we study in detail from approximately 1 billion to 219,855.

Figure 2 **Distribution of Cascade Sizes on a Log–Log Scale, Aggregated Across the Four Domains We Study: Videos, News, Pictures, and Petitions**



Note. CCDF = complementary cumulative distribution function.

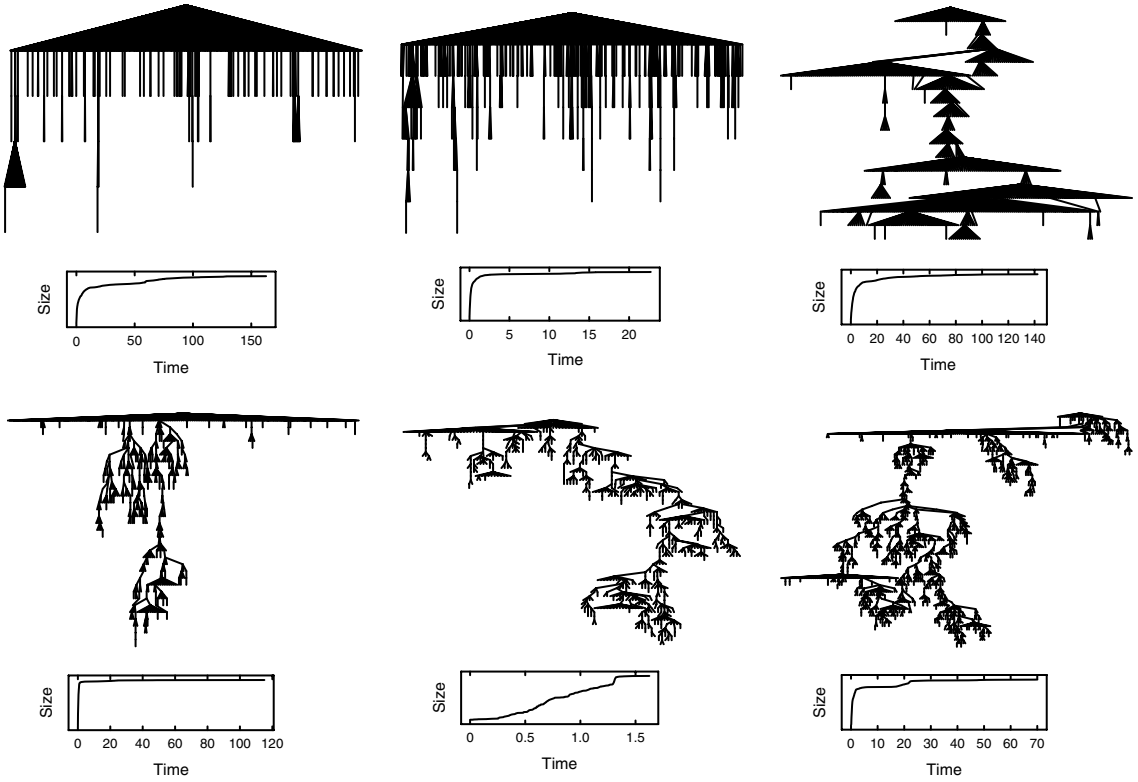
4.1. Structural Diversity

From this subpopulation of "successful" diffusion events, Figure 3 displays a stratified random sample ordered by structural virality $\nu(T)$. Specifically, cascades with between 100 and 1,000 adopters were ranked by $\nu(T)$ and logarithmically binned, and a random cascade was then drawn from each bin.[9] We

[8] Although diffusion trees are in reality dynamic objects, meaning that they grow over time as new adoptions take place, here we treat them as static objects representing the final state of a given diffusion process.

[9] We note that this exercise was performed only once to avoid hand selection of the best "random" sample.

Figure 3 A Random Sample of Cascades Stratified and Ordered by Increasing Structural Virality, Ranging from 2 to 50

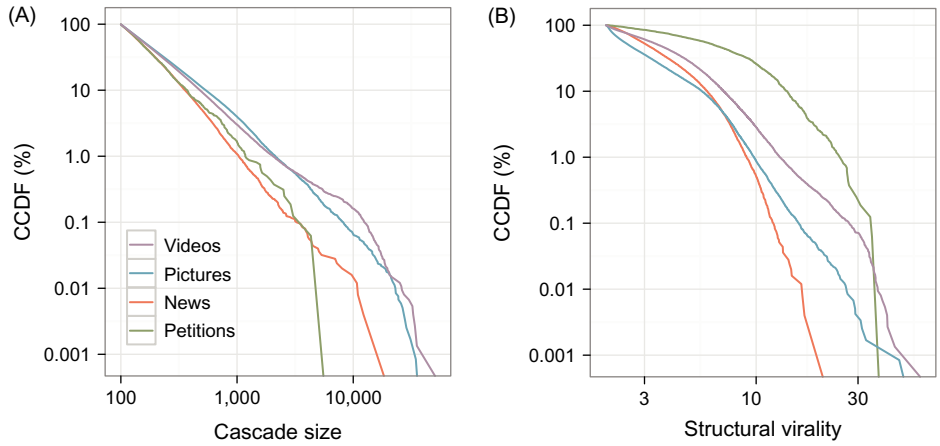


Notes. For ease of visualization, cascades were restricted to having between 100 and 1,000 adopters. Cumulative adoption curves (i.e., total cascade size over time) are shown below each cascade, with time indicated in hours. For visual clarity, the adoption curves terminate at 99% of the final cascade size.

observe that the ordering from left to right and top to bottom by increasing $\nu(T)$ is strikingly consistent with how these same structures would be ranked intuitively in order of increasing virality, not only in the trivial case of disambiguating broadcast and viral extremes, but also in making relatively fine-grained distinctions between intermediate cases. Thus, $\nu(T)$ not only seems to be a reasonable measure of structural virality in theory, but also performs well in practice. Considering now the cumulative adoption curves shown below each cascade in Figure 3, we make two further observations. First, although the shape of these adoption curves varies considerably, from events that experience a phase of rapid growth before leveling off to events that grow almost linearly over time, there is no consistent relationship with structural virality. Strikingly, in fact, the least structurally viral of all our sampled events (top left) exhibits a cumulative adoption curve that is almost indistinguishable in shape from the most structurally viral (bottom right). Second, the timescales on which the adoptions take place (noted in hours on the horizontal axis of the cumulative plots) also varies widely, from less than an hour (bottom left) to three days (top left). As with the shape of the curves, however, there is no consistent relationship between the timescale (speed) of an adoption process and its associated structural virality. We conclude that our measure of structural virality not only effectively quantifies differences in the underlying cascade structures, but is clearly doing so by using features of the diffusion process that are not captured by aggregated data.

The ordering also highlights our first main empirical finding: Although the structures in Figure 3 are all of similar size (i.e., have similar aggregate numbers of adopters), they exhibit remarkable diversity in structure, from an approximately pure broadcast ($\nu(T) \approx 2$, top left) to an ideal-type branching structure ($\nu(T) = 34$, bottom right), with numerous intermediate variations in between. The classical literature on diffusion often posits a critical threshold—or "tipping point"—for virality, suggesting a sharp break between cascades that are viral and those that are not. If the tipping point intuition is correct, one would expect that relatively large diffusion events such as those captured in the $n = 100$ (roughly one event in 4,000) to $n = 1{,}000$ (one in 100,000) range would be dominated either by broadcasts on the one hand or by viral spreading on the other hand, but that combinations of the two should not arise. More generally, one might expect only a handful of canonical forms to account for the majority of large events: for example, some events spread exclusively via broadcast, whereas others spread exclusively via word of

Figure 4 **Size and Structural Virality Distributions on a Log–Log Scale for Cascades Containing at Least 100 Adopters, Separated by Domain**



Note. CCDF, complementary cumulative distribution function.

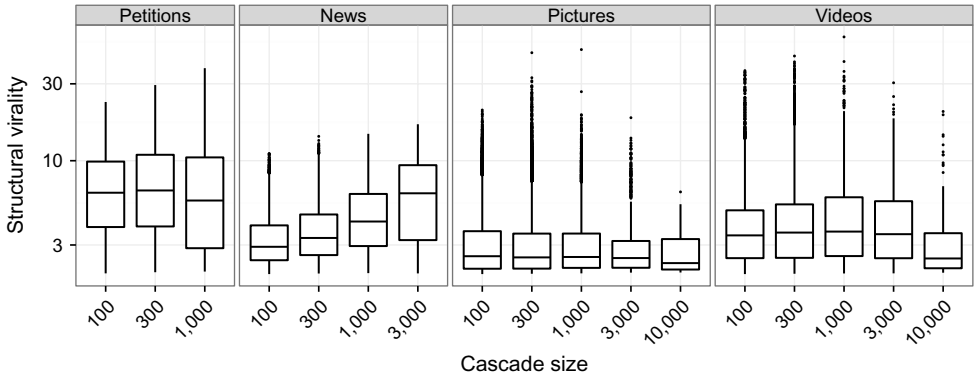
mouth, and others still spread by some combination of the two. In other words, whatever one's intuitive mental model of diffusion, one would likely expect to find that successful diffusion events of a given size would be typified by some combination of broadcast and viral diffusion, or at least some small taxonomy of types. It is striking, therefore, that Figure 3 shows examples of fine-grained variations in structural virality across the entire range of possibilities.

4.2. Examining Popularity and Structural Virality

Although Figure 3 shows that one can find examples of cascades across the spectrum of structural virality, it says little about their relative frequency or how that varies by domain. To address these questions, Figure 4(A) shows the size distribution of cascades larger than 100 adopters for all four domains—news, videos, images, and petitions—while Figure 4(B) shows the corresponding distributions of structural virality. As anticipated, Figure 4(A) shows that cascades can grow very large: For images and videos, the largest cascades attract several tens of thousands of reposts, whereas the most popular news stories are somewhat smaller (roughly 10,000 reposts), and petitions smaller still (several thousand reposts). In other words, although the vast majority of cascades are indeed small, large cascades do occur, albeit with low frequencies. Moreover, the size distributions appear to cluster into two categories: one comprising images and videos and the other comprising the rather less popular categories of petitions and news stories. In other words, the most popular videos and images are more popular than the most popular news stories and petitions not only because there are many more of the former, but also because the corresponding distributions exhibit a shallower slope; that is, for any given percentile of the relevant population, videos and images are more popular than petitions and news stories. Although we lack a compelling explanation for this systematic difference, we note that the vast majority of the most popular Twitter accounts belong not to news organizations or petition sites, but to celebrities, whose postings often contain images and videos. Moreover, YouTube and Instagram are among the top 10 most followed accounts, further facilitating the visibility of videos and images, respectively. It thus seems likely that one of the primary drivers of large image and video cascades is their promotion by individuals with large numbers of followers, consistent with past results (Bakshy et al. 2011).

Next, Figure 4(B) confirms the impression from Figure 3 that structural virality varies widely, from 2 (pure broadcast) to over 30. In particular, in contrast to classical "tipping point" theories of diffusion, we do not see a bimodal distribution of structural virality corresponding to broadcasts on the one hand and viral spreading on the other, but rather a continuous distribution of structural virality, confirming our earlier speculation that in some sense every conceivable combination of broadcasts and word-of-mouth transmission is represented. Interestingly, however, popular petitions are substantially more structurally viral than any other type of content, followed by videos, images, and news stories. For example, whereas about a quarter of popular petitions have structural virality of at least 10—meaning that petitions having garnered at least 100 adopters are quite likely to have grown virally—only about 3% of videos, 1% of images, and 0.5% of news stories exhibit the same level of structural virality. In spite of the diversity evident both in Figure 3 and Figure 4(B), therefore, the relatively larger size of cascades involving videos and images combined with their relatively low structural virality suggests that the largest cascades in those categories

Figure 5 **Box Plot of Structural Virality by Size on a Log–Log Scale, Separated by Domain**



Note. Lines inside the boxes indicate median structural virality, whereas the boxes themselves show interquartile ranges.

are not especially viral in a structural sense. In the next section, we examine this possibility in more detail.

4.3. Relationship Between Popularity and Structural Virality

As pointed out earlier, the relationship between popularity (cascade size) and structural virality is not a priori obvious; that is, depending on the empirically observed preponderance of broadcasts in small versus large events, the relationship could be positive (large events are less likely to be dominated by broadcasts than small events), negative (large events are more likely to be dominated by broadcasts than small events), or neither. Put another way, if cascades typically grow via person-to-person diffusion, we would expect structural virality to increase with cascade size. On the other hand, if large cascades are the product of broadcasts attributable to popular users on Twitter—the most popular of whom have tens of millions of followers—structural virality may not vary significantly with size, or could even decrease.

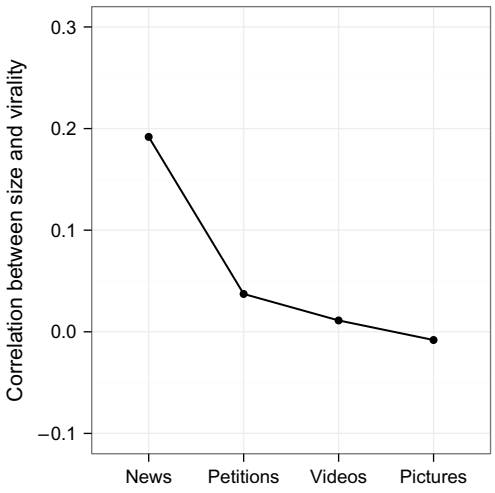
We investigate this question by examining the distribution of structural virality conditional on cascade size for each domain. First, and consistent with Figure 4, Figure 5 shows that across all sizes for which they occur, popular petitions are considerably more viral than the other domains. Second, Figure 5 shows that across all domains and size ranges, structural diversity varies considerably, confirming again the visual impression of Figure 3. Third, however, Figure 5 shows that for three out of four domains—petitions, images, and videos—median structural virality remains surprisingly invariant with respect to size. For images and videos, moreover, it is also surprisingly low: even the very largest cascades, comprising 10,000 reposts or more, exhibit median structural virality of less than 3, barely more than the theoretical minimum of 2. For petitions, meanwhile, median structural virality is between 7 and 8, roughly equivalent to a branching tree of depth between three and four generations: not a pure broadcast but still relatively shallow. Finally, for news, the relationship between size and structural virality is more positive than for the other domains, but also still surprisingly low. For cascades of size 100, for example, median structural virality is approximately 3, whereas for the largest observed news cascades, comprising 3,000 reposts, median structural virality is still less than 8, comparable to petitions.

We emphasize that there is nothing inevitable about this result. It could have been, for example, that the very largest events are characterized by multigenerational branching structures—indeed that is the clear implication of the phrase "going viral." So it is surprising that even the very largest events are, on average, dominated by broadcasts. It is also surprising that the correlation between size and structural virality is so low. As shown in Figure 6, the correlation for news is 0.2, indicating a positive but noisy relationship, whereas for petitions it is even lower (0.04), indicating almost no relationship at all, and for pictures and videos it is essentially zero. In contrast with our earlier result on diversity, which suggests that simply knowing the size of a cascade reveals very little about its structure, the combination of generally low values of structural virality and low correlation with size suggests that if popularity is consistently related to any one feature, it is the size of the largest broadcast.[10]

As in our discussion of Figure 4, we can only speculate about why (a) petitions are so much more structurally viral for every size category than other domains and (b) news stories show higher correlation between size and structural virality. We suspect, however, that the main driving factor is once again a relative dearth of large broadcast channels for petitions in particular and to a lesser extent news organizations.

[10] We also note that these results are not affected by the fact that the range of $\nu(T)$ varies with cascade size; the results are qualitatively identical when we use a measure of structural virality with a constant bounded range (see Appendix B).

Figure 6 **Correlation Between Cascade Size (Popularity) and Structural Virality Across Four Domains**



The popularity of images and videos, by contrast, is likely driven by celebrities, who increasingly have tens of millions of followers on Twitter, and whose posting behavior likely favors content of a personal and often visual nature over news and calls to action.

5. Theoretical Modeling

To recap, we have three main empirical findings. First, and consistent with previous work (Goel et al. 2012), the vast majority of diffusion events are small and accordingly lack much structure. Second, rare events that do become large exhibit striking structural diversity. And third, the size of these cascades is at most weakly correlated to their structural virality. Together these findings present an interesting theoretical question, namely, can they be replicated by a single underlying generative mechanism? And if so, what features are required? Although replicating some empirical results with a theoretical model does not on its own imply that the model is an accurate representation of the true generative process (Ijiri et al. 1977), it is nevertheless possible to rule some models out.

To address this question, we consider a series of variations on the SIR model, a classical model of biological contagion (Kermack and McKendrick 1927, Anderson and May 1991) that has frequently been adapted to model social diffusion processes,[11] initially to the specific context of new product adoption, where it is known as the Bass (1969) model, and subsequently to a wide range of other contexts including the propagation of links over a network of blogs (Leskovec et al. 2007). In any such model, there are two key sets of parameters. First, when an individual is infected (in the present case, with a piece of content), he or she subsequently infects each of his or her susceptible (i.e., not yet infected) contacts independently with probability β. Often β is assumed to be a constant, but in the current context—where it refers to the "infectiousness" of content—it is natural to think of it as being drawn from some distribution (which itself may be described by additional parameters). And second, we must specify the nature of the contact process, which here we model as a network in which $\bar{k}$ is the average node degree (i.e., the number of opportunities a typical node has to infect others) and σ^2 is the degree variance.[12]

Before proceeding, it is helpful to introduce the quantity $r = \bar{k}\beta$ (known in mathematical epidemiology as the "basic reproduction number" or R_0 of a disease). As alluded to earlier, a standard result for diseases spreading on random networks is that the condition $r = c$, where $c = 1/(1 + (\sigma/\bar{k})^2) \leq 1$, constitutes a critical threshold or tipping point, separating two regimes: a "supercritical," or "viral," regime $r > c$, in which small seeds can trigger exponential growth leading to large epidemics, and a "subcritical" regime $r < c$, in which the contagion almost surely dies out after infecting only a small number of susceptibles. From this general result, moreover, two more specific results follow. First, in Erdős–Rényi random networks $G(n, p)$, where the expected degree is $k \sim np$ and $\sigma^2 \sim k$ (as $n \to \infty$), the epidemic threshold condition reduces to $r \sim 1$ for $k \gg 1$. And second, in scale-free random networks (Barabási and Albert 1999) for which the variance diverges with the size of the network, it reduces to $r \sim 0$ as $n \to \infty$ (Pastor-Satorras and Vespignani 2001, Lyons 2000, Lloyd and May 2001), meaning that in sufficiently large scale-free networks, the subcritical regime effectively disappears.

These results are relevant to our analysis for two reasons. First, because viral events for which $r > 1$ exhibit exponential growth regardless of network structure and because we know from our data that large events are extremely rare, we restrict our analysis to the region $0 < r < 1$, corresponding to what in everyday usage would be thought of as "subcritical" spreading. Second, because we will consider both ER and scale-free random networks, the usual super-

[11] Reflecting its origins in mathematical epidemiology, the model is named for the three states—"susceptible," "infectious," and "recovered"—that each node in the network can occupy. Numerous variations of the basic SIR model have also been proposed, included the SI model, the SEIR model (where the "E" indicates "exposed"), the SIRS model, and so on (Anderson and May 1991). Here we refer to all such models canonically as SIR models.

[12] Additional parameters are also natural. For example, we only consider strict SIR models in the sense that after one time step, infected nodes are "removed" from the dynamics, meaning that they can no longer infect others nor become reinfected. Although natural for our case, where having "adopted" piece of content one cannot unadopt it, other assumptions are clearly possible, in which case additional parameters would be needed.

versus subcritical distinction is somewhat misleading. Specifically, whereas it does have a clear meaning for ER networks, for which only contagions with $r > 1$ are viral in the everyday sense of growing exponentially, in scale-free networks, all contagions are viral in the technical sense of exceeding the epidemic threshold, even though they are "dying out" as they attempt to spread.[13] As we will show next, in fact, models invoking ER networks are easily dismissed as incompatible with our empirical results, suggesting that the popular tipping point notion is largely irrelevant to the kind of viral events we study here.

We consider four models of increasing complexity and verisimilitude. In all cases, each realization of the simulation commences with an entirely susceptible population comprising 25 million individuals within which a single individual is randomly chosen to be the initially infected "seed" and proceeds until no further infections can take place.[14] We start by investigating contagions characterized by constant β spreading on an ER random graph. In light of the enormous attention paid to variations of this model both in the mathematical epidemiology (Kermack and McKendrick 1927, Anderson and May 1991) and marketing (Bass 1969, Valente 1995, Bass 2004) literatures, it is the natural baseline to consider. As noted above, however, its relevance to our empirical data can quickly be dismissed by showing that, consistent with standard theoretical results (Anderson and May 1991), the cascade size distribution is tightly centered around its mean regardless of the average network degree or infection rate, which is qualitatively different than the heavy-tailed size distribution we observe in the data.

One explanation for this result is that our assumption of constant β is unlikely to be correct. Presumably, content introduced to Twitter exhibits large differences in intrinsic interestingness and breadth of appeal, and therefore likelihood of being shared. This observation motivates the next model we consider, where the infection is again modeled as spreading on an ER graph, but the infectiousness of each piece of content, β_i, is now drawn from a power law distribution $\Pr(\beta_i) \sim \beta_i^{-\alpha}$, expressing the more plausible assumption that a large number of items in our sample are of low "quality" or "appeal" and hence are unlikely to spread (low β), whereas a small minority of appealing or high-quality items are much more likely to spread (high β). Studying this case, we do indeed recover the heavy-tailed size distribution from our empirical results. Interestingly, however, across parameter settings we consistently observe high correlation between cascade size and structural virality—because large cascades in ER must necessarily be multigenerational—which again stands in stark contrast to our empirical results. We therefore conclude that it is the ER network, not necessarily the assumption about constant item quality, that is responsible for the poor model fit.

Thus motivated, we now examine a third model in which we again assume β to be a constant, but the network is now a scale-free random network (Barabási and Albert 1999), constructed using the configuration method[15] (Newman 2005, Clauset et al. 2009), reflecting the roughly power law degree distribution $p(k) \sim k^{-\alpha}$ observed for Twitter (Bakshy et al. 2011). Sweeping over the two parameters, α and β, we simulated content of varying infectiousness diffusing over networks with varying degree skew. Figure 7 shows the results of nearly 100 billion simulations, with 1 billion cascades generated for each parameter setting (α, β), roughly congruent with the number of cascades we analyzed on Twitter. Figure 7 shows that for certain parameters—$r \approx 0.5$ and $\alpha \approx 2.3$—the model recapitulates several important features of our empirical data.[16] First, Figure 7(A) shows that for this parameter setting the probability of a given piece of content becoming "popular"—meaning that it attracts at least 100 adoptions—is consistent with the observed rate of roughly one in one thousand. Second, Figure 7(B) shows that the mean structural virality for these parameters is 5, which again is in line with our observations. Third, Figure 7(C) shows that the correlation between size and structural virality is also in the observed range. Finally, Figure 8 shows the full marginal distributions of size and virality, and the distribution of virality conditional on

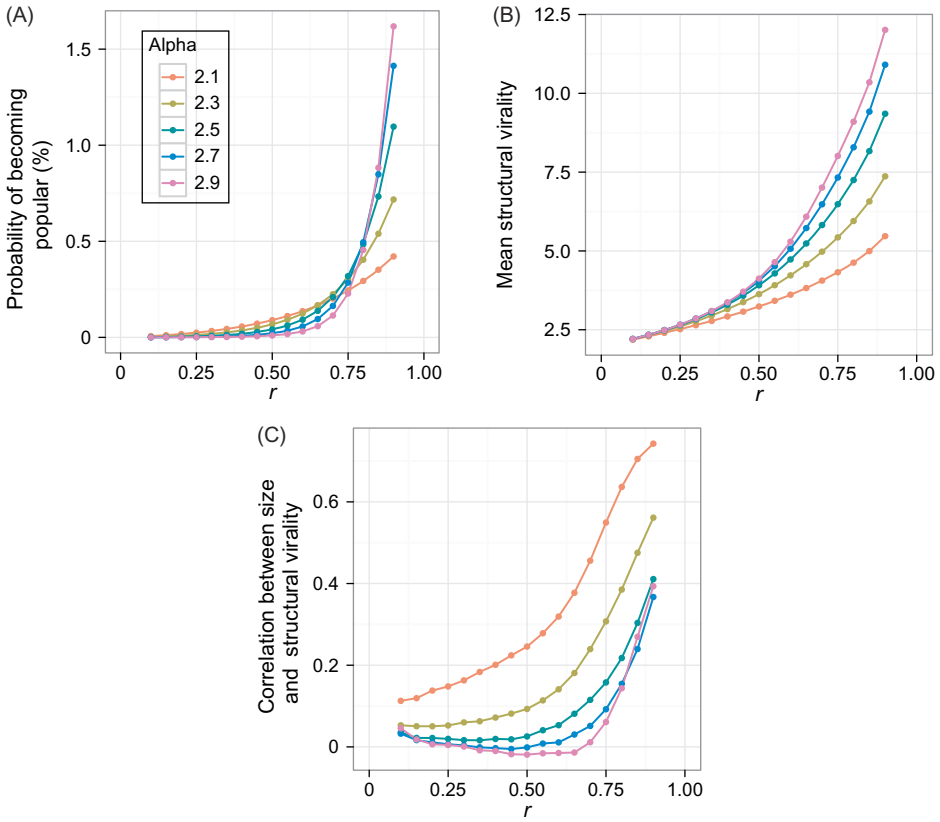
[13] The intuitive explanation for this counterintuitive result is that in scale-free networks, a typical node is likely to be connected via at most a short path to a "hub" node with an extremely high degree that, if infected, can sustain an infection that would ordinarily die out (Pastor-Satorras and Vespignani 2001).

[14] Clearly on Twitter a single unique piece of content can be introduced many times independently. In such cases, there is potential for two cascades to "collide," which clearly cannot happen in our simulations, where we introduce only one seed at a time. In light of the extreme rarity of large cascades, however, and the large size of the Twitter network, such collisions are also rare; hence, we do not believe this simplification has any significant consequences. We also note that our model is a special case of what has been called "simple contagion" (Centola 2010), in which the infection probably is independent across multiple exposures. In contrast with "complex contagion," such as occurs in "threshold models" (Granovetter 1978), where multiple exposures can combine in highly nonlinear ways, the use of individual seeds for simple contagion is relatively unproblematic.

[15] For each node in the network, its number of followers (i.e., out-degree) was first randomly selected according to a discrete power law degree distribution with exponent α, a minimum value of 10, and a maximum value of 1 million. Then nodes in the networks were randomly connected while preserving the specified degrees.

[16] The power law exponent of $\alpha \approx 2.3$ is consistent with the observed degree distribution on Twitter (Kwak et al. 2010).

Figure 7 **Likelihood of Becoming Popular (i.e., Having at Least 100 Adopters), Mean Structural Virality, and the Correlation Between Size and Structural Virality for Simulated Cascades Generated from an SIR Model on a Random Scale-Free Network, Plotted as a Function of the Model Parameters**



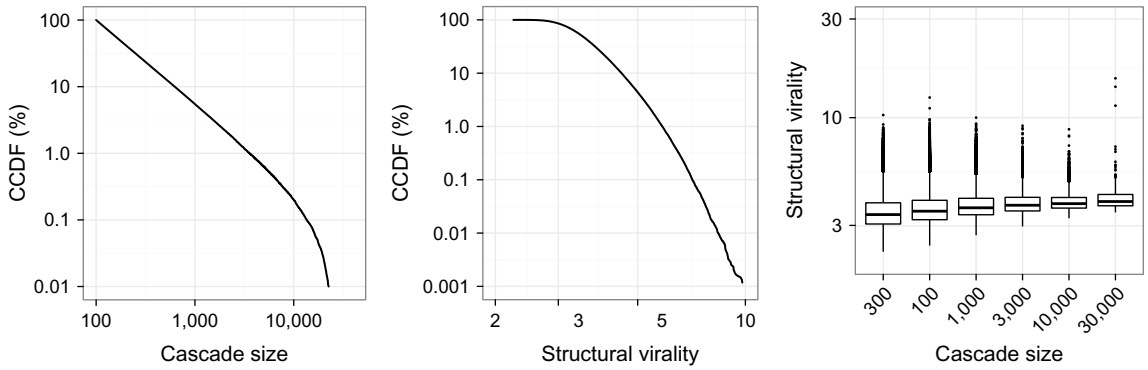
Note. Each line corresponds to a different exponent α for the power-law network degree distribution, and $r = \beta\bar{k}$ is the expected number of individuals a random node infects in a fully susceptible population.

size for this parameter choice, where we again see that the simulated cascades are similar to the empirically observed events. One notable difference between empirical and simulation results, however, is that the variance in each bin (as measured by the interquartile range) in the rightmost plot in Figure 8 is considerably less than that in Figure 5, indicating that empirical cascades exhibit much more structural diversity at any given size compared to those generated by the model.

These simulation results can be interpreted in two ways. On the one hand, it is striking that so simple a model—with only two tunable parameters—can capture many of the basic empirical regularities of what is undoubtedly a far more complex and multifaceted system. For example, although the success of real-world products is almost certainly affected by their quality, this connection is absent from our model. Indeed, for any fixed parameter choice under the SIR model, all cascades—the largest broadcasts, the most viral cascades, and the many events that acquire only a handful of adopters—have the same infectiousness β. In other words, taking infectiousness as a proxy for quality, in our simulations the largest and most viral cascades are not inherently better than those that fail to gain traction, but are simply more fortunate (Watts 2002). On the other hand, it is also interesting that our model is not able to fully capture the diversity of structural virality exhibited in the empirical data. Although we can only speculate on the reasons for this limitation, two possible explanations immediately suggest themselves. The simplest explanation is that as large as our simulated networks are (25 million nodes), they are still not as large nor is the network structure as complex as the actual Twitter follower graph, which comprises roughly 500 million users, the most connected of whom have well over 50 million followers. Possibly, therefore, the difference could be accounted for simply by increasing the size of the networks by another one or two orders of magnitude—an increase that is computationally challenging, but that is straightforward in theory. A second, and perhaps more likely, explanation is that our assumption of constant β remains too simplistic, and that introducing such variation into our model would also increase the variation of structural virality at any given size.

The fourth and final model that we simulate therefore replaces constant β with β_i drawn from a power

Figure 8 **Box Plot of Structural Virality by Size (on a Log–Log Scale) for 1 Billion Simulated Cascades Generated from an SIR Model on a Random Scale-Free Network with $\alpha = 2.3$ and $r = 0.5$**



Note. CCDF, complementary cumulative distribution function.

law distribution, identical to the ER case in our second model above. Surprisingly, however, a similarly extensive set of simulations using this model finds that it does not in fact lead to noticeably more structural diversity; moreover, it leads to high correlation between size and structural virality. The reason for both results is that higher (lower) values of β_i generate larger (smaller) events, not more (less) structurally viral events of the same size. Thus, even though the diversity of β_i does affect the size distribution of cascades, for a given cascade size it does not generate more diversity of structural virality. Identifying a mechanism that accounts for the observed diversity of structural virality therefore presents an interesting challenge for future modeling work.

6. Discussion

Returning to our opening motivation, our paper makes three main contributions. First, we have introduced the concept of structural virality, one of the first measures to formally quantify the structure of information cascades. Although our results are restricted to the diffusion of information on Twitter, our structural approach to diffusion processes applies quite generally, both to online and offline settings. It is often claimed, for example, that some of the most successful Internet products in recent history, such as Hotmail, Gmail, and Facebook, were driven primarily by word-of-mouth adoption, in part because the companies that created these products did not initially have large advertising budgets, and in part because by design they contained features to explicitly encourage sharing. Yet these products also benefitted from extensive media coverage, which might have driven large numbers of adoptions from a small number of broadcast events. Likewise, although popular Internet memes are typically described as having spread virally, they also typically receive substantial media coverage. Without reconstructing the actual sequence of events by which a given product, idea, or piece of content was adopted, and relatedly without a metric for quantifying virality, the mere observation of popularity—however rapidly accrued—allows one to conclude little about the relative importance of viral versus broadcast mechanisms in determining the observed outcome. With the appropriate data, therefore, our notion of structural virality could conceivably shed light on a much broader range of diffusion processes than we have considered here.

Our second contribution is to measure the fine-grain structure of nearly 1 billion naturally occurring diffusion events in a specific online setting, namely, Web content spreading on Twitter. In particular, we have identified hundreds of thousands of large cascades—the biggest such collection to date—revealing remarkable structural diversity of diffusion events, ranging from broadcast to viral and containing essentially everything in between, where we emphasize that such an exercise would be difficult absent a metric for classifying and ordering the structure of these cascades automatically. In addition, we find relatively low correlation between size and virality, highlighting the difficulty of determining how content spread given only knowledge of its popularity.

Third, we have shown that a simple model of contagion is broadly consistent with our empirical findings. The theoretical literature has largely focused on supercritical diffusion processes to model large, viral cascades; however, the vast majority of diffusion events comprise only a few nodes, and rarely extend beyond one generation beyond the root node, or seed (Goel et al. 2012). Events of this latter kind are naturally attributable to subcritical diffusion,[17] and hence one might thus be tempted to model online diffusion via two categorically distinct mechanisms, separately accounting for the head and tail

[17] For example, Leskovec et al. (2007) found that a susceptible-infected-susceptible (SIS) model with $\beta = 0.025$, equivalent to $r \approx 0.14$, was able to replicate the size distribution of observed cascades of links over a network of blogs.

of the distribution. Indeed, the very label "viral hit" implies precisely the exponential spreading of the sort observed in contagion models in their supercritical regime. It is therefore notable that essentially everything we observe, including the very largest and rarest events, can be accounted for by a simple model operating entirely in the low infectiousness parameter regime. Indeed our best model fit is for $r \approx 0.5$, which is considerably lower even than a previous "subcritical" estimate of $\beta \approx 0.99$ based on the diffusion of chain letters (Golub and Jackson 2010)—a difference that is likely due to the heavy-tailed (scale-free) degree distribution of Twitter.[18]

Finally, in addition to our three scientific contributions, we note that our work also contributes to the emerging field of computational social science in the sense that it addresses a traditional social science question—How does content spread via social networks?—but answers it using a type and scale of data that has only recently become available; that is, only after tracing the propagation of over a billion pieces of content can we collect an unbiased sample of large, and exceedingly rare, cascades to observe their subtle structural properties. By contrast, previous work (Goel et al. 2012) that investigated the propagation of nearly one million news stories and videos—one of the largest diffusion studies at the time—was only able to observe relatively small events, resulting in a qualitatively incomplete view of diffusion. In a similar vein, the most relevant previous analysis of the structure of extremely large diffusion events relied on just two examples, specifically the reconstructed paths of two Internet chain letters (Liben-Nowell and Kleinberg 2008). Although collecting even two such examples required considerable ingenuity, it is nevertheless the case that inferring general principles from so few observations is inherently difficult (Golub and Jackson 2010, Chierichetti et al. 2011). One of our main findings, in fact, is that large diffusion events exhibit extreme diversity of structural forms—a finding that necessarily requires many examples. Thus, although our current work is by no means exhaustive, its scale facilitates a significant step toward describing the nature and diversity of online information diffusion.

Appendix A. Computing Structural Virality

The average distance measure of structural virality that we use, $\nu(T)$, has often been applied in mathematical chemistry, where it is known as the Wiener index, and its efficient computation has also long been known. For completeness, here we present a simple and scalable method to compute $\nu(T)$. We begin by showing how the Wiener index, as well as the average depth of a tree, can be expressed in terms of the sizes of various subtrees.

Lemma 1. *For a tree T with n nodes, let $\text{depth}_{\text{avg}}$ denote the average depth of nodes in the tree. Letting $\mathcal{S}$ be the set of all subtrees of T, we have*

$$\frac{1}{n}\sum_{S\in\mathcal{S}}|S| = \text{depth}_{\text{avg}} + 1.$$

Proof. For any node $v_i \in T$ and any subtree $S \in \mathcal{S}$, let $\delta_S(v_i)$ be 1 if $v_i \in S$ and 0 otherwise. Then,

$$\begin{aligned}\sum_{S\in\mathcal{S}}|S| &= \sum_{S\in\mathcal{S}}\sum_{i=1}^{n}\delta_S(v_i)\\ &= \sum_{i=1}^{n}\sum_{S\in\mathcal{S}}\delta_S(v_i)\\ &= \sum_{i=1}^{n}1+\text{depth}(v_i).\end{aligned}$$

The result now follows by dividing each side by n. □

Theorem 2. *For a tree T with n nodes, let $\text{depth}_{\text{avg}}$ denote the average depth of nodes in the tree, let dist_{avg} denote the average distance between all pairs of distinct nodes (i.e., $\text{dist}_{\text{avg}} = \nu(T)$), and let $\mathcal{S}$ be the set of all subtrees of T. Then,*

$$\text{dist}_{\text{avg}} = \frac{2n}{n-1}\left[1+\text{depth}_{\text{avg}} - \frac{1}{n^2}\sum_{S\in\mathcal{S}}|S|^2\right]. \tag{A1}$$

In particular,

$$\text{dist}_{\text{avg}} = \frac{2n}{n-1}\left[\frac{1}{n}\sum_{S\in\mathcal{S}}|S| - \frac{1}{n^2}\sum_{S\in\mathcal{S}}|S|^2\right]. \tag{A2}$$

Proof. Statement (A2) in the theorem follows directly from (A1) together with Lemma 1, and so we only need to establish statement (A1). For any two nodes $v_i, v_j \in T$, let $\text{LCA}(v_i, v_j)$ denote their lowest common ancestor: the unique node in T of greatest depth that has both v_i and v_j as descendants (where a node is allowed to be a descendant of itself). Since the shortest path between v_i and v_j goes through $\text{LCA}(v_i, v_j)$, we have

$$\begin{aligned}\text{dist}(v_i, v_j) &= \text{dist}(v_i, \text{LCA}(v_i, v_j)) + \text{dist}(\text{LCA}(v_i, v_j), v_j)\\ &= [\text{depth}(v_i) - \text{depth}(\text{LCA}(v_i, v_j))]\\ &\quad + [\text{depth}(v_j) - \text{depth}(\text{LCA}(v_i, v_j))]\\ &= \text{depth}(v_i) + \text{depth}(v_j) - 2\cdot\text{depth}(\text{LCA}(v_i, v_j)).\end{aligned}$$

Let $\text{subtrees}(v_i, v_j)$ be the set of subtrees that contain both v_i and v_j, and observe that this set consists of exactly those subtrees that contain $\text{LCA}(v_i, v_j)$. Since for any node v there are $1 + \text{depth}(v)$ subtrees that contain it,

$$|\text{subtrees}(v_i, v_j)| = 1 + \text{depth}(\text{LCA}(v_i, v_j)).$$

Substituting this expression into the previous equation, we see that

$$\text{dist}(v_i, v_j) = 2 + \text{depth}(v_i) + \text{depth}(v_j) - 2|\text{subtrees}(v_i, v_j)|.$$

[18] We note that this finding also recalls earlier work that sought to account for the surprisingly long-term and low-level persistence of computer viruses in terms of a low-infectiousness contagion spreading over a scale-free network (Pastor-Satorras and Vespignani 2001). Although that work did not address the structural properties of the events in question, the mechanism identified as responsible—namely, low-infectiousness contagion combined with the occasional encounter with a high-degree node—is largely similar to the one investigated here.

For any node $v_i \in T$ and any subtree $S \in \mathcal{S}$, let $\delta_S(v_i)$ be 1 if $v_i \in S$ and 0 otherwise. Then, summing over all n^2 pairs of nodes, we have

$$\sum_{i,j=1}^{n} \text{dist}(v_i, v_j) = 2n^2 + 2n\sum_{i=1}^{n} \text{depth}(v_i) - 2\sum_{i,j=1}^{n}\sum_{S\in\mathcal{S}} \delta_S(v_i)\delta_S(v_j)$$

$$= 2n^2 + 2n\sum_{i=1}^{n} \text{depth}(v_i) - 2\sum_{S\in\mathcal{S}} |S|^2.$$

The result follows by dividing through by $n(n-1)$ the number of pairs of distinct nodes. □

Theorem 2 shows that $\nu(T)$ can be expressed in terms of the sizes of subtrees of T. Algorithm 1 uses this observation to efficiently compute $\nu(T)$.

Algorithm 1 (Computing $\nu(T)$)
Require: T is a tree rooted at node r

```
function SUBTREE-MOMENTS(T, r)
    if T.size() = 1 then                          ▷ The base case
        size ← 1
        sum-sizes ← 1
        sum-sizes-sqr ← 1
    else                ▷ Recurse over the children of the root r
        for c ∈ r.children() do
            size_c, sum-sizes_c, sum-sizes-sqr_c
                ← SUBTREE-MOMENTS(T, c)
        size ← 0
        sum-sizes ← 0
        sum-sizes-sqr ← 0
        for c ∈ r.children() do
            size ← size + size_c
            sum-sizes ← sum-sizes + sum-sizes_c
            sum-sizes-sqr ← sum-sizes-sqr
                                + sum-sizes-sqr_c
        size ← size + 1
        sum-sizes ← sum-sizes + size
        sum-sizes-sqr ← sum-sizes-sqr + size^2
    return size, sum-sizes, sum-sizes-sqr
function AVERAGE-DISTANCE(T, r)
    size, sum-sizes, sum-sizes-sqr
        ← SUBTREE-MOMENTS(T, r)
    dist_avg ← [2 · size/(size − 1)]×
                [sum-sizes/size − sum-sizes-sqr/size^2]
    return dist_avg
```

Table B.1 Rank Correlation Between Alternative Measures of Structural Virality

	Average distance	Relative broadcast	Distinct parent	Average depth
Average distance	1	−0.79	0.73	0.90
Relative broadcast	−0.79	1	−0.98	−0.66
Distinct parent	0.73	−0.98	1	0.61
Average depth	0.90	−0.66	0.61	1

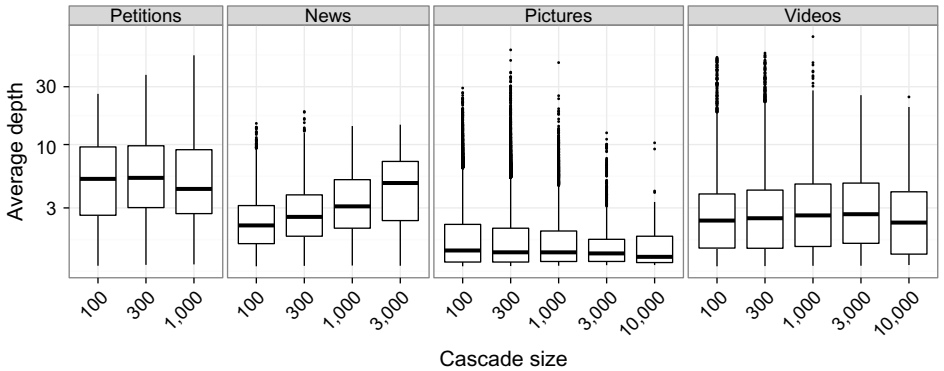
Appendix B. Alternative Measures of Structural Virality

Although we have demonstrated that our particular definition of structural virality is reasonable, there are several other formalizations of the concept that also qualify as reasonable candidates. In particular, here we consider the following three metrics:

1. the relative size of the largest broadcast (i.e., the largest number of children of any single node in the diffusion tree, as a fraction of the total number of nodes in the tree);
2. the probability that two randomly selected nodes have a distinct parent node;
3. the average depth of nodes in the tree.

Simple inspection shows that all three of these alternatives distinguish between the extremes of a single, large broadcast on the one hand and a multigenerational "viral" cascade on the other. However, they all capture subtly different structural aspects of diffusion trees, and also fail for somewhat different pathological cases. Consequently, as with our primary definition above, it is difficult to evaluate the utility of the various metrics on theoretical grounds alone, or even to assess their similarity. In practice, however, we find that they are all highly correlated with our chosen average path length measure, and also with each other. Specifically, Table B.1 shows that when computed over the entire set of empirically observed cascades with at least 100 adopters, $\nu(T)$ has an absolute rank correlation greater than 0.73 with all three alternative measures. Moreover, our empirical results are qualitatively similar regardless of which of these alternative measures of structural virality we apply. For example, Figure B.1 shows the relationship between size and average depth, analogous to Figure 5, and from which essentially the same conclusions could be drawn.

Figure B.1 Box Plot of an Alternative Measure of Structural Virality—Average Cascade Depth—by Size (on a Log Scale), Separated by Domain



Note. Lines inside the boxes indicate the medians, whereas the boxes themselves show interquartile ranges.

Thus, although we cannot rule out the possibility that a superior metric to ours can be defined, we can at least substantiate two related claims: first, that our choice of metric is at least roughly as good as a number of other plausible candidates, and second, that our substantive findings are robust with respect to the particular manner in which we formalize the concept of structural virality.

Appendix C. Tree Construction Method

Here we describe the process of constructing a diffusion tree for a particular piece of content (e.g., a given URL). Trees are composed of one node for each user who has adopted the content, and each edge links a user back to an inferred "parent." After each adoption has been identified as either a root or the child of another post, we construct the cascade of adoptions.

In an ideal setting we would have access to this information for each adoption, but in practice these details are not always available. The best-case scenario is use of Twitter's official retweet functionality, which enables a user to effectively forward a tweet that was originally authored by someone else. Attribution is clear in these cases, and tree construction would be relatively straightforward if all adoptions were of this form. Unfortunately, however, users also repost content using a variety of unofficial conventions, which complicate the attribution task. For instance, the unofficial retweet convention amounts to copying the text of a tweet and prepending "RT @username" to credit another individual. Twitter treats these posts as originally authored content and has no formal way of linking them back to original posts. Finally, users may forego crediting a source entirely, in which case one must make an educated guess about who (if anyone) in their feed exposed them to the content and who should be credited as responsible for their adoption.

We decompose the process of inferring a parent into two steps, described in detail below. We estimate that our inference procedure correctly identifies the parent of an adoption in approximately 95% of instances.

1. *Identify potential parents.* For each user who adopts a piece of content, we identify a set of "potential parents," defined as individuals whose adoption of a piece of content appears in the focal user's timeline prior to the focal user's adoption. In other words, potential parents are the set of individuals who are likely to have exposed the user to the adopted content. To identify these potential parents, we note that a user's timeline contains (1) all posts originally authored by the user's friends and (2) tweets authored by others that at least one of the user's friends has "officially retweeted" using Twitter's built-in reposting functionality. In particular, any tweet appears at most once in a user's timeline regardless of how many of his or her friends have officially retweeted it.[19] To compute the set of potential parents for a given adoption, we join activity from the Twitter Firehose application programming interface (API), which provides details about each tweet, with the Twitter follower graph, which provides the listing of who follows whom.

[19] Any nonofficial reposting—e.g., using the "RT @username" convention—is considered originally authored, resulting in potentially repeated content in a user's timeline.

2. *Infer a single parent.* We now identify the single most likely parent from the set of all potential parents of a given adoption. To do this, we consider three cases based on how the focal user posted the content.

a. *Official retweet.* If the focal user officially retweeted a post that appeared in their timeline (i.e., retweeted the post via Twitter's built-in functionality), then the Twitter API provides the ID of the original tweet. We then use this information to identify the individual who introduced the post to the user's timeline as the parent. We note that the parent need not be the original author of the tweet—for example, in the case of a friend who retweeted a third party, as described above. Also, users occasionally officially retweet content that did not appear in their timelines (e.g., because they discovered it by browsing); in these cases we treat the focal user as a "root" and do not assign a parent. Overall, in these official retweet cases—which constitute 65% of the instances we consider—we almost certainly correctly attribute the tweet.

b. *Accredited repost.* In the case of a nonofficial retweet, credit may still be present in the form of a mentioned user, for example, using the "RT @username" convention. We identify as the parent the individual who most recently introduced a post of that content, authored by the mentioned user, to the focal user's timeline. This mentioned user may be a friend of the focal user, in which case the friend is assigned as the parent. Alternatively, the mentioned user may be a third party—e.g., a friend of a friend. In this case, the friend who most recently mentioned the accredited user along side the piece of content is identified as the parent. As above, if no such friend can be identified, we treat the focal user as a root and do not assign a parent. Accredited posts constitute 10% of the adoptions we analyze, and as in the case of official retweets, the inferred parent is almost certainly correct.

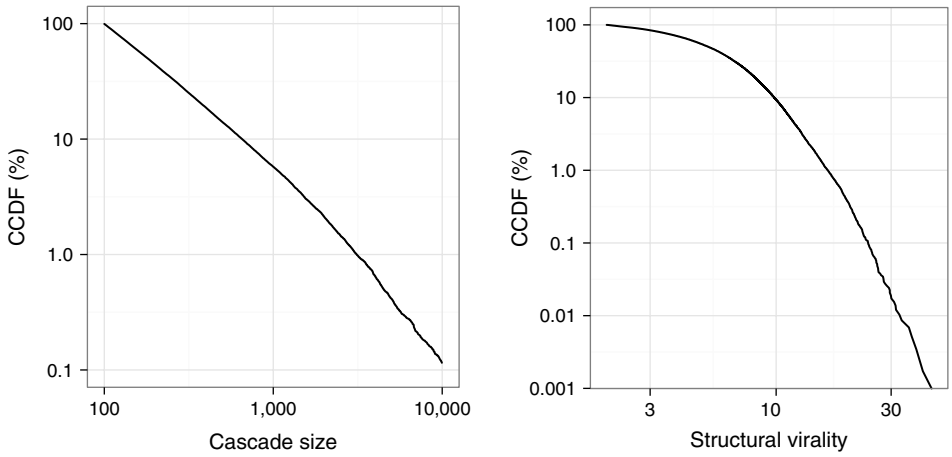
c. *Uncredited repost.* In this final, case we lack any explicit information about how the user was exposed to the content and simply assign as the parent the friend who most recently introduced the content to the focal user's timeline. If no such friend exists, we again treat the focal user as a root. To assess the accuracy of our inference strategy in this case, we apply it to the set of official retweets, for which we are fairly certain which individual is the parent of any given adoption. We find that the most-recent-introduction heuristic correctly identifies the parent 79% of the time.

Since our inference procedure almost certainly identifies the correct parent in the first two cases—official retweets and accredited reposts, which together account for 75% of adoptions—and since we estimate 79% accuracy for the remaining 25% of adoptions, we conclude that the overall accuracy of our parent inference strategy is 95%.

Appendix D. Off-Channel Diffusion

Although our empirical findings are qualitatively quite similar across the four distinct domains studied above, it is possible that all four suffer from one of two systematic biases that might affect our conclusions. First, a potential problem with studying the diffusion of external content on Twitter (e.g., news stories from the *New York Times* and videos from YouTube) is that the same content may also spread via other channels, such as Facebook or email. As a result of this "off-channel" diffusion, two individuals on Twitter who appear

Figure D.1 Size and Structural Virality Distributions on a Log–Log Scale for Popular Hashtag Cascades, Containing at Least 100 Adopters

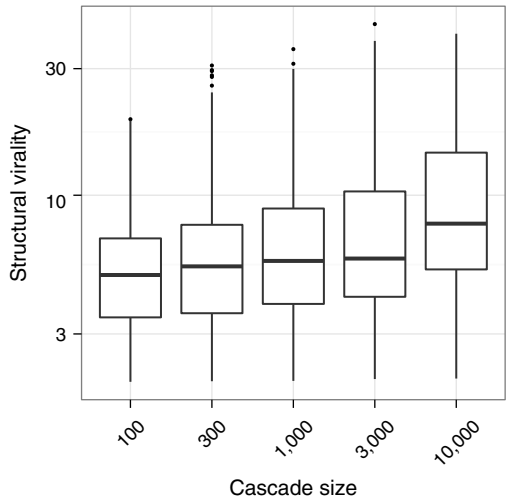


Note. CCDF, complementary cumulative distribution function.

to have introduced the same piece of content independently may in fact be connected, thus leading us to mistakenly treat a single diffusion tree as two disjoint events. A second concern is that our use of reposting rather than retweeting also potentially biases our data. Specifically, user–follower similarity (i.e., homophily) may lead connected users to post the same content independently in close temporal sequence, leading us to conflate similarity with influence (Shalizi and Thomas 2011, Aral et al. 2009, Lyons 2011).

To check that off-channel diffusion does not systematically bias our findings, we consider the diffusion of Twitter-specific "hashtags"—short fragments of text used to indicate the topic of a tweet. Because such hashtags are less likely to have originated outside of Twitter, and because for the same reason they are less likely to migrate off of Twitter, these data are correspondingly less susceptible to any biases associated with off-channel diffusion. Moreover, to ensure as much as possible that we are considering only on-Twitter uses of hashtags, we restrict our sample to "long" hashtags, which are especially unlikely to be used elsewhere. To define "long," we note that hashtags on Twitter are generally written in camel case (e.g., #CamelCase). Treating each substring that begins with a capitalized letter and ends immediately before the next capitalized letter as a "word," we trace the diffusion of hashtags that include five or more such words (e.g., #ThisIsALongHashtag). As infrequent as these long hashtags are relative to hashtags in general, they are still plentiful, amounting to 58,000 cascades with at least 100 adopters. Figures D.1 and D.2 show that the diffusion of these long hashtags yields qualitatively similar results to our primary analysis, suggesting that off-channel diffusion is not driving our findings.

Figure D.2 Box Plot of Structural Virality by Size on a Log–Log Scale for Hashtag Cascades



Note. Lines inside the boxes indicate the median structural virality, whereas the boxes themselves show interquartile ranges.

References

Adar E, Adamic LA (2005) Tracking information epidemics in blogspace. *IEEE/WIC/ACM Internat. Conf. Web Intelligence* (Institute of Electrical and Electronics Engineers, Piscataway, NJ).

Anderson RM, May RM (1991) *Infectious Diseases of Humans* (Oxford University Press, Oxford, UK).

Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA* 106(51):21544–21549.

Bakshy E, Karrer B, Adamic LA (2009) Social influence and the diffusion of user-created content. *Proc. Tenth ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 325–334.

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: Quantifying influence on twitter. *Proc. Fourth ACM Internat. Conf. Web Search and Data Mining* (Association for Computing Machinery, New York), 65–74.

Barabási AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.

Bass FM (1969) A new product growth for model consumer durables. *Management Sci.* 15(5):215–227.

Bass FM (2004) Comments on "a new product growth for model consumer durables the bass model." *Management Sci.* 50(12 supplement):1833–1840.

Centola D (2010) The spread of behavior in an online social network experiment. *Science* 329(5996):1194–1197.

Chierichetti F, Kleinberg J, Liben-Nowell D (2011) Reconstructing patterns of information diffusion from incomplete observations. *Adv. Neural Inform. Processing Systems*, Vol. 24 (Neural Information Processing Systems Foundation, La Jolla, CA).

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. *SIAM Rev.* 51(4):661–703.

Coleman J, Katz E, Menzel H (1957) The diffusion of an innovation among physicians. *Sociometry* 20(4):253–270.

Dodds PS, Watts DJ (2004) Universal behavior in a generalized model of contagion. *Phys. Rev. Lett.* 92(21):218701.

Dow PA, Adamic LA, Friggeri A (2013) The anatomy of large Facebook cascades. *Proc. Seventh Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Fichman RG (1992) Information technology diffusion: A review of empirical research. *Proc. 13th Internat. Conf. Inform. Systems* (University of Minnesota, Minneapolis), 195–206.

Goel S, Watts DJ, Goldstein DG (2012) The structure of online diffusion networks. *Proc. 13th ACM Conf. Electronic Commerce* (Association for Computing Machinery, New York), 623–638.

Golub B, Jackson MO (2010) Using selection bias to explain the observed structure of Internet diffusions. *Proc. Natl. Acad. Sci. USA* 107(24):10833–10836.

Granovetter M (1978) Threshold models of collective behavior. *Amer. J. Sociol.* 83(6):1420–1443.

Hoang T-A, Lim E-P (2012) Virality and susceptibility in information diffusions. *Proc. Sixth Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Ijiri Y, Simon HA, Bonini CP, van Wormer TA (1977) *Skew Distributions and the Sizes of Business Firms* (North-Holland Publishing Company, New York).

Iyengar R, Van den Bulte C, Valente TW (2010) Opinion leadership and social contagion in new product diffusion. *Marketing Sci.* 30(2):195–212.

Jenders M, Kasneci G, Naumann F (2013) Analyzing and predicting viral tweets. *Proc. 22nd Internat. Conf. World Wide Web Companion* (International World Wide Web Conferences Steering Committee, New York), 657–664.

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. *Proc. Ninth ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York), 137–146.

Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proc. Roy. Soc. Lond. A* 115(772): 700–721.

Kupavskii A, Ostroumova L, Umnov A, Usachev S, Serdyukov P, Gusev G, Kustarev A (2012) Prediction of retweet cascade size over time. *Proc. 21st ACM Internat. Conf. Inform. Knowledge Management* (Association for Computing Machinery, New York), 2335–2338.

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? *WWW '10: Proc. 19th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 591–600.

Leskovec J, Adamic LA, Huberman BA (2007) The dynamics of viral marketing. *ACM Trans. Web* 1(1):5.

Leskovec J, Singh A, Kleinberg J (2006) Patterns of influence in a recommendation network. *Adv. Knowledge Discovery Data Mining* 3918:380–389.

Liben-Nowell D, Kleinberg J (2008) Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA* 105(12):4633–4638.

Lloyd AL, May RM (2001) How viruses spread among computers and people. *Science* 292(5520):1316–1317.

Lopez-Pintado D, Watts DJ (2008) Social influence, binary decisions and collective dynamics. *Rationality Soc.* 20(4):399–443.

Lyons R (2000) Phase transitions on nonamenable graphs. *J. Math. Phys.* 41(3):1099–1126.

Lyons R (2011) The spread of evidence-poor medicine via flawed social-network analysis. *Statist., Politics, Policy* 2(1).

Ma Z, Sun A, Cong G (2013) On predicting the popularity of newly emerging hashtags in Twitter. *J. Amer. Soc. Inform. Sci. Tech.* 64(7):1399–1410.

Mahajan V, Peterson RA (1985) *Models for Innovation Diffusion*, Vol. 48 (Sage, Newbury Park, CA).

Mohar B, Pisanski T (1988) How to compute the wiener index of a graph. *J. Math. Chemistry* 2(3):267–277.

Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. *Contemporary Phys.* 46(5):323–351.

Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86(14):3200–3203.

Petrovic S, Osborne M, Lavrenko V (2011) RT to win! Predicting message propagation in Twitter. *Proc. Fifth Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Rogers EM (1962) *Diffusion of Innovations* (Free Press, New York).

Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* 40(2):211–239.

Sun E, Rosenn I, Marlow C, Lento T (2009) Gesundheit! Modeling contagion through facebook news feed. *Proc. Third Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA).

Toole JL, Cha M, González MC (2012) Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS One* 7(1):e29528.

Tsur O, Rappoport A (2012) What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. *Proc. Fifth ACM Internat. Conf. Web Search and Data Mining* (Association for Computing Machinery, New York), 643–652.

Valente TW (1995) *Network Models of the Diffusion of Innovations*, Quantitative Methods in Communication Series (Hampton Press, Cresskill, NJ).

Van den Bulte C, Lilien GL (2001) Medical innovation revisited: Social contagion versus marketing effort1. *Amer. J. Sociol.* 106(5):1409–1435.

Walther JB, Carr CT, Choi SSW, DeAndrea DC, Kim J, Tong ST, Van Der Heide B (2010) Interaction of interpersonal, peer, and media influence sources online. *A Networked Self: Identity, Community, and Culture on Social Network Sites*, Vol. 17 (Routledge, London).

Watts DJ (2002) A simple model of information cascades on random networks. *Proc. Natl. Acad. Sci. USA* 99(9):5766–5771.

Wiener H (1947) Structural determination of paraffin boiling points. *J. Amer. Chemical Soc.* 69(1):17–20.

Wu S, Hofman JM, Mason WA, Watts DJ (2011) Who says what to whom on Twitter. *Proc. 20th Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 705–714.

Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in Twitter. *Proc. Fourth Internat. AAAI Conf. Weblogs Soc. Media* (AAAI Press, Palo Alto, CA), 355–358.

Yang J, Leskovec J (2010) Modeling information diffusion in implicit networks. *Proc. 10th IEEE Internat. Conf. Data Mining* (IEEE Computer Society, Washington, DC), 599–608.

Young PH (2009) Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *Amer. Econom. Rev.* 99(5):1899–1924.

An Experimental Study of the Small World Problem

Jeffrey Travers; Stanley Milgram

Sociometry, Vol. 32, No. 4 (Dec., 1969), 425-443.

Stable URL:
http://links.jstor.org/sici?sici=0038-0431%28196912%2932%3A4%3C425%3AAESOTS%3E2.0.CO%3B2-W

Sociometry is currently published by American Sociological Association.

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

AND

STANLEY MILGRAM

The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target person with fewer intermediaries than those starting in Nebraska; subpopulations in the Nebraska group do not differ among themselves. The funneling of chains through sociometric "stars" is noted, with 48 per cent of the chains passing through three persons before reaching the target. Applications of the method to studies of large scale social structure are discussed.

The simplest way of formulating the small world problem is "what is the probability that any two people, selected arbitrarily from a large population, such as that of the United States, will know each other?" A more interesting formulation, however, takes account of the fact that, while persons a and z may not know each other directly, they may share one or more mutual acquaintances; that is, there may exist a set of individuals, B, (consisting of individuals b_1, b_2 . . . b_n) who know both a and z and thus link them to one another. More generally, a and z may be connected not by any single common acquaintance, but by a series of such intermediaries, a-b-c- . . . -y-z; i.e., a knows b (and no one else in the chain); b knows a and in addition knows c, c in turn knows d, etc.

To elaborate the problem somewhat further, let us represent the popula-

* The study was carried out while both authors were at Harvard University, and was financed by grants from the Milton Fund and from the Harvard Laboratory of Social Relations. Mr. Joseph Gerver provided invaluable assistance in summarizing and criticizing the mathematical work discussed in this paper.

tion of the United States by a partially connected set of points. Let each point represent a person, and let a line connecting two points signify that the two individuals know each other. (Knowing is here assumed to be symmetric: if *a* knows *b* then *b* knows *a*. Substantively, "knowing" is used to denote a mutual relationship; other senses of the verb, e.g. knowing about a famous person, are excluded.) The structure takes the form of a cluster of roughly 200 million points with a complex web of connections among them. The acquaintance chains described above appear as pathways along connected line segments. Unless some portion of the population is totally isolated from the rest, such that no one in that subgroup knows anyone outside it, there must be at least one chain connecting any two people in the population. In general there will be many such pathways, of various lengths, between any two individuals.

In view of such a structure, one way of refining our statement of the small world problem is the following: given two individuals selected randomly from the population, what is the probability that the minimum number of intermediaries required to link them is 0, 1, 2, . . . *k*? (Alternatively, one might ask not about the minimum chains between pairs of people, but mean chain lengths, median chain lengths, etc.)

Perhaps the most direct way of attacking the small world problem is to trace a number of real acquaintance chains in a large population. This is the technique of the study reported in this paper. The phrase "small world" suggests that social networks are in some sense tightly woven, full of unexpected strands linking individuals seemingly far removed from one another in physical or social space. The principal question of the present investigation was whether such interconnectedness could be demonstrated experimentally.

The only example of mathematical treatment dealing directly with the small world problem is the model provided by Ithiel Pool and Manfred Kochen (unpublished manuscript). Pool and Kochen assume a population of N individuals, each of whom knows, on the average, n others in the population. They attempt to calculate P_k, the probability that two persons chosen randomly from the group can be linked by a chain of k intermediaries. Their basic model takes the form of a "tree" or geometric progression. Using an estimate of average acquaintance volume provided by Gurevitch (1961), they deduce that two intermediaries will be required to link typical pairs of individuals in a population of 200 million. Their model does not take account of social structure. Instead of allowing acquaintance nets to define the boundaries of functioning social groups, Pool and Kochen must, for the purposes of their model, conceive of society as being partitioned into a number of hypothetical groups, each with identical populations. They are then able

to devise a way to predict chain lengths within and between such hypothesized groups.

In an empirical study related to the small world problem Rapoport and Horvath (1961) examined sociometric nets in a junior high school of 861 students. The authors asked students to name in order their eight best friends within the school. They then traced the acquaintance chains created by the students' choices. Rapoport was interested in connectivity, i.e. the fraction of the total population that would be contacted by tracing friendship choices from an arbitrary starting population of nine individuals. Rapoport and his associates (Rapoport and Horvath, 1961; Foster et al., 1963; Rapoport, 1953; 1963) have developed a mathematical model to describe this tracing procedure. The model takes as a point of departure random nets constructed in the following manner: a small number of points is chosen from a larger population and a fixed number of "axones" is extended from each of these points to a set of target points chosen at random from the population. The same fixed number of axones is then extended from each of the target points to a set of second generation target points, and the process is repeated indefinitely. A target point is said to be of the tth remove if it is of the tth generation and no lower generation. Rapoport then suggests a formula for calculating the fraction, P_t, of the population points which are targets of the tth remove. He is also able to extend the formula to nonrandom nets, such as those created in the Rapoport and Horvath empirical study, by introducing a number of "biases" into the random net model. Rapoport shows that two parameters, obtainable from the data, are sufficient to produce a close fit between the predictions of the model and the empirical outcome of the trace procedure.[1]

Rapoport's model was designed to describe a trace procedure quite different from the one employed in the present study; however, it has some relation to the small world problem. If we set the number of axones traced from a given individual equal to the total number of acquaintances of an average person, the Rapoport model predicts the total fraction of the population potentially traceable at each remove from the start, serving precisely the aims of the model of Pool and Kochen. (It should, however, be noted that Rapoport's model deals with asymmetric nets, and it would be difficult to modify the model to deal with general symmetric nets, which characterize the small world phenomenon.)

Despite the goodness of fit between Rapoport's model and the data from

[1] There is additional empirical evidence (Fararo and Sunshine, 1964) and theoretical support (Abelson, 1967) for the assumption that two parameters are sufficient to describe the Rapoport tracing procedure, i.e. that more complex biases have minimal effects on connectivity in friendship nets.

two large sociograms, there are unsolved problems in the model, as Rapoport himself and others (Fararo and Sunshine, 1964) have pointed out. The Pool-Kochen model involves assumptions difficult for an empirically oriented social scientist to accept, such as the assumption that society may be partitioned into a set of groups alike in size and in internal and external connectedness. In the absence of empirical data, it is difficult to know which simplifying assumptions are likely to be fruitful. On the other hand, with regard to the empirical study of Rapoport and Horvath, the fact that the total population employed was small, well-defined, and homogeneous leaves open many questions about the nature of acquaintance nets in the larger society.[2] An empirical study of American society as a whole may well uncover phenomena of interest both in their own right and as constraints on the nature of any correct mathematical model of the structure of large-scale acquaintanceship nets.

PROCEDURE

This paper follows the procedure for tracing acquaintance chains devised and first tested by Milgram (1967). The present paper introduces an experimental variation in this procedure, by varying "starting populations"; it also constitutes a first technical report on the small world method.

The procedure may be summarized as follows: an arbitrary "target person" and a group of "starting persons" were selected, and an attempt was made to generate an acquaintance chain from each starter to the target. Each starter was provided with a document and asked to begin moving it by mail toward the target The document described the study, named the target, and asked the recipient to become a participant by sending the document on. It was stipulated that the document could be sent only to a first-name acquaintance of the sender. The sender was urged to choose the recipient in such a way as to advance the progress of the document toward the target; several items of information about the target were provided to guide each new sender in his choice of recipient. Thus, each document made its way along an acquaintance chain of indefinite length, a chain which would end only when it reached the target or when someone along the way declined to participate. Certain basic information, such as age, sex and occupation, was collected for each participant.

[2] In addition to the Pool-Kochen and Rapoport work, there are numerous other studies of social network phenomena tangentially related to the small-world problem. Two well-known examples are Bailey's *The Mathematical Theory of Epidemics* and Coleman, Katz and Menzel's *Medical Innovation.* Bailey's work deals with diffusion from a structured source, rather than with convergence on a target from a set of scattered sources, as in the present study. The Coleman, Katz and Menzel study deals with an important substantive correlate of acquaintance nets, namely information diffusion.

We were interested in discovering some of the internal structural features of chains and in making comparisons across chains as well. Among the questions we hoped to answer were the following: How many of the starters —if any—would be able to establish contact with the target through a chain of acquaintances? How many intermediaries would be required to link the ends of the chains? What form would the distribution of chain lengths take? What degree of homogeneity in age, sex, occupation, and other characteristics of participants would be observed within chains? How would complete chains differ from incomplete on these and other dimensions?

An additional comparison was set up by using three distinct starting subpopulations. The target person was a Boston stockbroker; two of the starting populations were geographically removed from him, selected from the state of Nebraska. A third population was selected from the Boston area. One of the Nebraska groups consisted of bluechip stockholders, while the second Nebraska group and the Boston group were "randomly" selected and had no special access to the investment business. By comparisons across these groups we hoped to assess the relative effects of geographical distance and of contact with the target's occupational group. Moreover we hoped to establish a strategy for future experimental extensions of the procedure, in which the sociological characteristics of the starting and target populations would be systematically varied in order to expose features of social structure.

The primary research questions, then, involved a test of the feasibility and fruitfulness of the method as well as an attempt to discover some elementary features of real social nets. Several experimental extensions of the procedure are already underway. A more detailed description of the current method is given in the following sections.

PARTICIPANTS. *Starting Population.* The starting population for the study was comprised of 296 volunteers. Of these, 196 were residents of the state of Nebraska, solicited by mail. Within this group, 100 were systematically chosen owners of blue-chip stocks; these will be designated "Nebraska stockholders" throughout this paper. The rest were chosen from the population at large; these will be termed the "Nebraska random" group. In addition to the two Nebraska groups, 100 volunteers were solicited through an advertisement in a Boston newspaper (the "Boston random" group). Each member of the starting population became the first link in a chain of acquaintances directed at the target person.

Intermediaries. The remaining participants in the study, who numbered 453 in all, were in effect solicited by other participants; they were acquaintances selected by previous participants as people likely to extend the chain toward the target. Participation was voluntary. Participants were not paid, nor was money or other reward offered as incentive for completion of chains.

THE DOCUMENT. The 296 initial volunteers were sent a document which was the principal tool of the investigation.[3] The document contained:

a. a description of the study, a request that the recipient become a participant, and a set of rules for participation;
b. the name of the target person and selected information concerning him;
c. a roster, to which each participant was asked to affix his name;
d. a stack of fifteen business reply cards asking information about each participant.

Rules for Participation. The document contained the following specific instructions to participants:

a. Add your name to the roster so that the next person who receives this folder will know whom it came from.
b. Detach one postcard from the bottom of this folder. Fill it out and return it to Harvard University. No stamp is needed. The postcard is very important. It allows us to keep track of the progress of the folder as it moves toward the target person.
c. If you know the target person on a personal basis, mail this folder directly to him (her). Do this only if you have previously met the target person and know each other on a first name basis.
d. If you do not know the target person on a personal basis, do not try to contact him directly. Instead, mail this folder to a personal acquaintance who is more likely than you to know the target person. You may send the booklet on to a friend, relative, or acquaintance, but it must be someone you know personally.

Target Person. The target person was a stockholder who lives in Sharon, Massachusetts, a suburb of Boston, and who works in Boston proper. In addition to his name, address, occupation and place of employment, participants were told his college and year of graduation, his military service dates, and his wife's maiden name and hometown. One question under investigation was the type of information which people would use in reaching the target.

Roster. The primary function of the roster was to prevent "looping," i.e., to prevent people from sending the document to someone who had already received it and sent it on. An additional function of the roster was to motivate people to continue the chains. It was hoped that a list of prior participants, including a personal acquaintance who had sent the document to

[3] A photographic reproduction of this experimental document appears in Milgram, 1969: 110–11.

the recipient, would create willingness on the part of those who received the document to send it on.

Tracer Cards. Each participant was asked to return to us a business reply card giving certain information about himself and about the person to whom he sent the document. The name, address, age sex and occupation of the sender and sender's spouse were requested, as were the name, address, sex and age of the recipient. In addition, the nature of the relationship between sender and recipient—whether they were friends, relatives, business associates, etc.—was asked. Finally, participants were asked why they had selected the particular recipient of the folder.

The business reply cards enabled us to keep running track of the progress of each chain. Moreover, they assured us of getting information even from chains which were not completed, allowing us to make comparisons between complete and incomplete chains.

RESULTS

COMPLETIONS. 217 of the 296 starting persons actually sent the document on to friends. Any one of the documents could reach the target person only if the following conditions were met: 1) recipients were sufficiently motivated to send the document on to the next link in the chain; 2) participants were able to adopt some strategy for moving the documents closer to the target (this condition further required that the given information allow them to select the next recipient in a manner that increased the probability of contacting the target); 3) relatively short paths were in fact required to link starters and target (otherwise few chains would remain active long enough to reach completion). Given these contingencies, there was serious doubt in the mind of the investigators whether any of the documents, particularly those starting in an area remote from the target person, could move through interlocking acquaintance networks and converge on him. The actual outcome was that 64 of the folders, or 29 per cent of those sent out by starting persons, eventually reached the target.

DISTRIBUTION OF CHAIN LENGTHS. *Complete Chains.* Figure 1 shows the frequency distribution of lengths of the completed chains. "Chain length" is here defined as the number of intermediaries required to link starters and target. The mean of the distribution is 5.2 links.

It was unclear on first inspection whether the apparent drop in frequency at the median length of five links was a statistical accident, or whether the distribution was actually bimodal. Further investigation revealed that the summary relation graphed in Figure 1 concealed two underlying distributions: when the completed chains were divided into those which approached the target through his hometown and those which approached him via

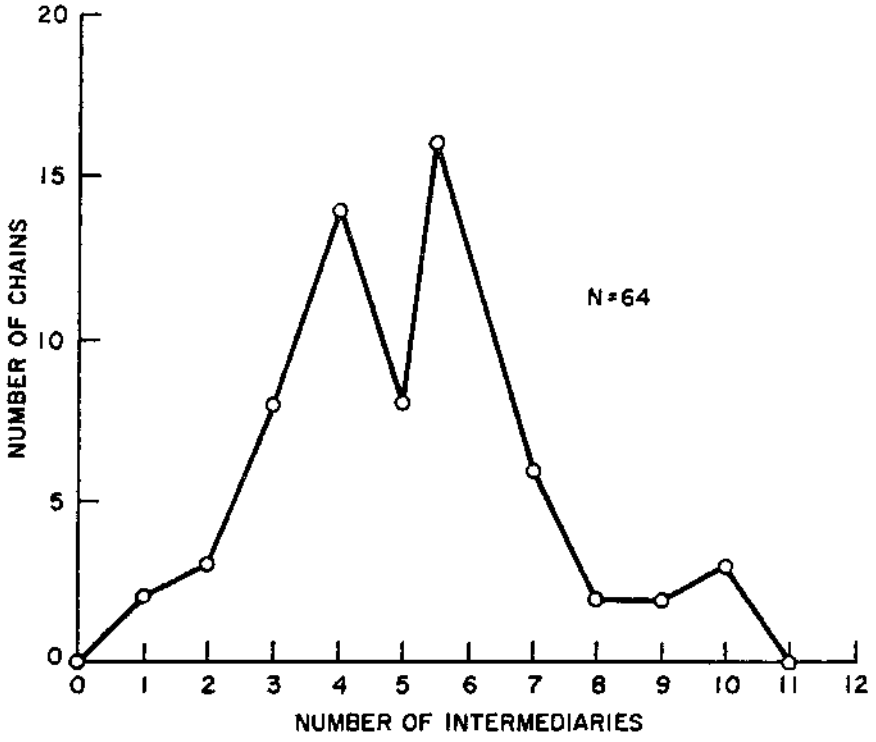


FIGURE 1

Lengths of Completed Chains

Boston business contracts, two distinguishable distributions emerged. The mean of the Sharon distribution is 6.1 links, and that of the Boston distribution is 4.6. The difference is significant at a level better than .0005, as assessed by the distribution-free Mann-Whitney *U* test. (Note that more powerful statistical tests of the significance of differences between means cannot be applied to these data, since those tests assume normality of underlying distributions. The shape of the true or theoretical distribution of lengths of acquaintance chains is precisely what we do not know.)

Qualitatively, what seems to occur is this. Chains which converge on the target principally by using geographic information reach his hometown or the surrounding areas readily, but once there often circulate before entering the target's circle of acquaintances. There is no available information to narrow the field of potential contacts which an individual might have within the town. Such additional information as a list of local organizations

of which the target is a member might have provided a natural funnel, facilitating the progress of the document from town to target person. By contrast, those chains which approach the target through occupational channels can take advantage of just such a funnel, zeroing in on him first through the brokerage business, then through his firm.

Incomplete Chains. Chains terminate either through completion or dropout: each dropout results in an incomplete chain. Figure 2 shows the number of chains which dropped out at each "remove" from the starting population. The "0th remove" represents the starting population itself: the "first remove" designates the set of people who received the document directly from members of the starting population. The "second remove" received the document from the starters via one intermediary, the third through two intermediaries, etc. The length of an incomplete chain may be defined as the number of removes from the start at which dropout occurs, or, equivalently, as the number of transmissions of the folder which precede dropout. By this definition, Figure 2 represents a frequency distribution of the lengths of incomplete chains. The mean of the distribution is 2.6 links.

The proportion of chains which drop out at each remove declines as

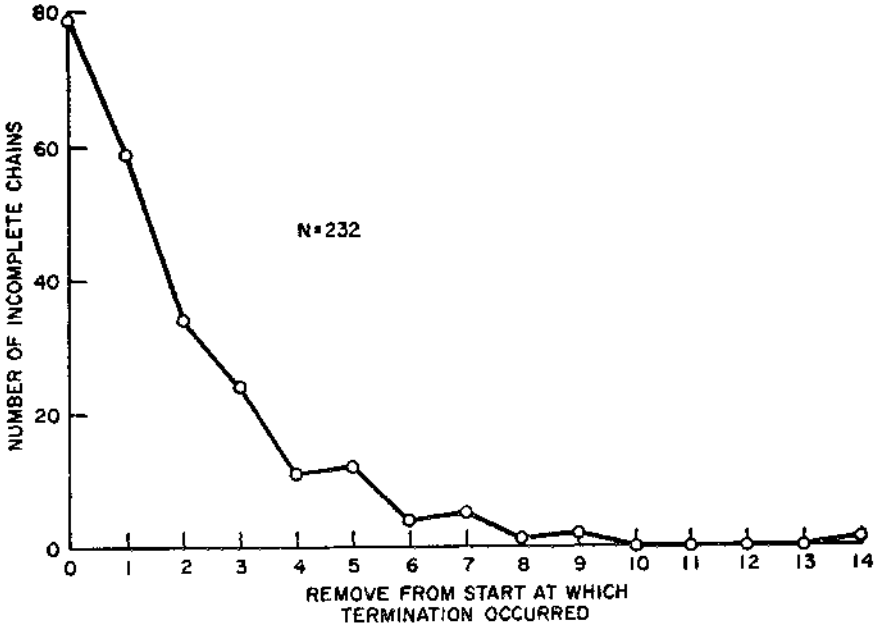


FIGURE 2

Lengths of Incomplete Chains

chains grow in length, if that proportion is based on all chains active at each remove (those destined for completion as well as incompletion). About 27 per cent of the 296 folders sent to the starting population are not sent on. Similarly, 27 per cent of the 217 chains actually initiated by the starters die at the first remove. The percentage of dropouts then appears to fall. It also begins to fluctuate, as the total number of chains in circulation grows small, and an increasing proportion of completions further complicates the picture.

It was argued earlier that, in theory, any two people can be linked by at least one acquaintance chain of finite length, barring the existence of totally isolated cliques within the population under study. Yet, incomplete chains are found in our empirical tracing procedure because a certain proportion of those who receive the document do not send it on. It is likely that this occurs for one of two major reasons: 1) individuals are not motivated to participate in the study; 2) they do not know to whom to send the document in order to advance it toward the target.

For purposes of gauging the significance of our numerical results, it would be useful to know whether the dropouts are random or systematic, i.e., whether or not they are related to a chain's prognosis for rapid completion. It seems possible, for example, that dropouts are precisely those people who are least likely to be able to advance the document toward the target. If so, the distribution of actual lengths of completed chains would understate the true social distance between starters and target by an unknown amount. (Even if dropouts are random, the observed distribution understates the true distribution, but by a potentially calculable amount.) We can offer some evidence, however, that this effect is not powerful.

First, it should be clear that, though people may drop out because they see little possibility that any of their acquaintances can advance the folder toward the target, their subjective estimates are irrelevant to the question just raised. Such subjective estimates may account for individual decisions not to participate; they do not tell us whether chains that die in fact would have been longer than others had they gone to completion. People have poor intuitions concerning the lengths of acquaintance chains. Moreover, people can rarely see beyond their own acquaintances; it is hard to guess the circles in which friends of friends—not to mention people even more remotely connected to oneself—may move.

More direct evidence that dropouts may be treated as "random" can be gleaned from the tracer cards. It will be recalled that each participant was asked for information not only about himself but also about the person to whom he sent the document. Thus some data were available even for dropouts, namely age, sex, the nature of their relationship to the people

TABLE 1

Activity of Chains at Each Remove

All Chains					Incomplete Chains Only			
Remove	Chains Reaching this Remove	Completions at this Remove	Dropouts at this Remove	Per cent Dropouts	Remove	Chains Reaching this Remove	Dropouts at this Remove	Per cent Dropouts
0	296	0	79	27	0	232	79	34
1	217	0	59	27	1	153	59	39
2	158	2	34	22	2	94	34	36
3	122	3	24	20	3	60	24	40
4	95	8	11	12	4	36	11	31
5	76	14	12	16	5	25	12	48
6	50	8	4	8	6	13	4	31
7	38	16	5	13	7	9	5	55
8	17	6	1	6	8	4	1	25
9	10	2	2	20	9	3	2	67
10	6	2	0	0	10	1	0	0
11	4	3	0	0	11	1	0	0
12	1	0	0	0	12	1	0	0
13	1	0	0	0	13	1	0	0
14	1	0	1	100	14	1	1	100

preceding them in the chain, and the reason the dropout had been selected to receive the document. These four variables were tabulated for dropouts versus non-dropouts. None of the resulting contingency tables achieved the .05 level of statistical significance by chi-square test; we are therefore led to accept the null hypothesis of no difference between the two groups, at least on this limited set of variables. Of course, a definitive answer to the question of whether dropouts are really random must wait until the determinants of chain length are understood, or until a way is found to force all chains to completion.[4]

SUBPOPULATION COMPARISONS. A possible paradigm for future research using the tracing procedure described here involves systematic variation of the relationship between the starting and target populations. One such study, using Negro and White starting and target groups, has already been completed by Korte and Milgram (in press). In the present study, which involved only a single target person, three starting populations were used (Nebraska random, Nebraska stockholders, and Boston random.) The relevant experimental questions were whether the proportion of completed chains or mean chain lengths would vary as a function of starting population.

Chain Length. Letters from the Nebraska subpopulations had to cover a geographic distance of about 1300 miles in order to reach the target, whereas letters originating in the Boston group almost all started within 25 miles of his home and/or place of work. Since social proximity depends in part on geographic proximity, one might readily predict that complete chains originating in the Boston area would be shorter than those originating in Nebraska. This presumption was confirmed by the data. As Table 2 shows, chains originating with the Boston random group showed a mean length of 4.4 intermediaries between starters and target, as opposed to a mean length of 5.7 intermediaries for the Nebraska random group. ($p \leqq .001$ by

[4] Professor Harrison White of Harvard University has developed a technique for adjusting raw chain length data to take account of the dropout problem. His method assumes that dropouts are "random," in the following sense. An intermediary who knows the target sends him the folder, completing the chain, with probability 1. Otherwise, an intermediary throws away the folder with fixed probability $1-a$, or sends it on with probability a. If sent on, there is a probability Q_i (which depends on number of removes from the origin) that the next intermediary knows the target. The data is consistent with a value for a of approximately 0.75, independent of remove from the origin, and hence with a "random" dropout rate of 25 per cent. The limited data further suggest that Q_i grows in a "staircase" pattern from zero (at zero removes from the starting population) to approximately one-third at six removes, remaining constant thereafter. Based on these values, the hypothetical curve of completions with no dropouts resembles the observed curve shifted upward; the median length of completed chains rises from 5 to 7, but no substantial alteration is required in conclusions drawn from the raw data.

TABLE 2

Lengths of Completed Chains

Frequency Distribution

Population	Number of Intermediaries												
	0	1	2	3	4	5	6	7	8	9	10	11	Total
Nebraska Random	0	0	0	1	4	3	6	2	0	1	1	0	18
Nebraska Stock	0	0	0	3	6	4	6	2	1	1	1	0	24
Boston Random	0	2	3	4	4	1	4	2	1	0	1	0	22
All	0	2	3	8	14	8	16	6	2	2	3	0	64

Means

Starting Population	Mean Chain Length
Nebraska Random	5.7
Nebraska Stockholders	5.4
All Nebraska	5.5
Boston Random	4.4
All	5.2

a one-tailed Mann-Whitney U test.) Chain length thus proved sensitive to one demographic variable—place of residence of starters and target.

The Nebraska stockholder group was presumed to have easy access to contacts in the brokerage business. Because the target person was a stockbroker, chains originating in this group were expected to reach the target more efficiently than chains from the Nebraska random group. The chain-length means for the two groups, 5.7 intermediaries for the random sample and 5.4 for the stockholders, differed in the expected direction, but the difference was not statistically significant by the Mann-Whitney test. The stockholders used the brokerage business as a communication channel more often than did the random group; 60.7 per cent of all the participants in chains originating with the stockholder group reported occupations connected with finance, while 31.8 per cent of participants in chains originating in the Nebraska random group were so classified.

Proportion of Completions. As indicated in Table 3, the proportions of chains completed for the Nebraska random, Nebraska stockholder, and Boston subpopulations were 24 per cent, 31 per cent and 35 per cent, respectively. Although the differences are not statistically significant, there is a weak tendency for higher completion rates to occur in groups where mean length of completed chains is shorter. This result deserves brief discussion.

Let us assume that the dropout rate is constant at each remove from the start. If, for example, the dropout rate were 25 per cent then any chain would have a 75 per cent probability of reaching one link, $(.75)^2$ of reaching two links, etc. Thus, the longer a chain needed to be in order to reach completion, the less likely that the chain would survive long enough to run its full course. In this case, however, chain-length differences among the three groups were not sufficiently large to produce significant differences in completion rate. Moreover, if the dropout rate declines as chains grow long, such a decrease would off-set the effect just discussed and weaken the observed inverse relation between chain length and proportion of completions.

TABLE 3

Proportion of Completions for Three Starting Populations

	Starting Population							
	Nebraska Random		Nebraska Stock.		Boston		Total	
Complete	18	(24%)	24	(31%)	22	(35%)	64	(29%)
Incomplete	58	(76%)	54	(69%)	41	(65%)	153	(71%)
	76	(100%)	78	(100%)	63	(100%)	217	(100%)

$\chi^2=2.17$, df.$=2$, $p>.3$, N.S.

COMMON CHANNELS. As chains converge on the target, common channels appear—that is, some intermediaries appear in more than one chain. Figure 3 shows the pattern of convergence. The 64 letters which reached the target were sent by a total of 26 people. Sixteen, fully 25 per cent, reached the target through a single neighbor. Another 10 made contact through a single business associate, and 5 through a second business associate. These three "penultimate links" together accounted for 48 per cent of the total completions. Among the three, an interesting division of labor appears. Mr. G,

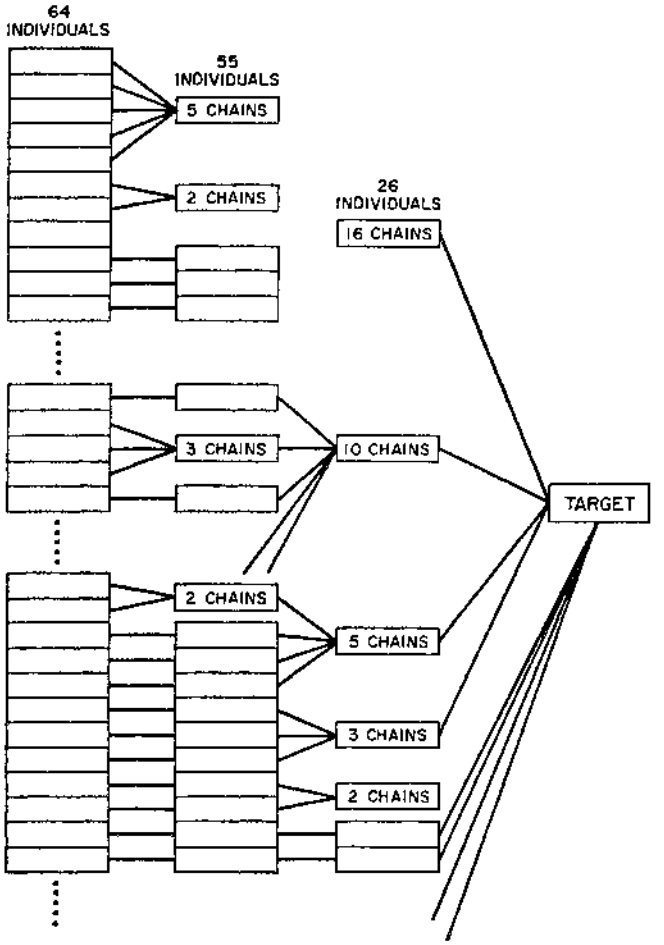


FIGURE 3

Common Paths Appear as Chains Converge on the Target

who accounted for 16 completions, is a clothing merchant in the target's hometown of Sharon; Mr. G funnelled toward the target those chains which were advancing on the basis of the target's place of residence. Twenty-four chains reached the target from his hometown; Mr. G accounted for ⅔ of those completions. All the letters which reached Mr. G came from residents of Sharon. By contrast, Mr. D and Mr. P, who accounted for 10 and 5 completions, respectively, were contacted by people scattered around the Boston area, and in several cases, by people living in other cities entirely. On the other hand, whereas Mr. G received the folder from Sharon residents in a wide variety of occupations, D and P received it almost always from stockbrokers. A scattering of names appear two or three times on the list of penultimate links; seventeen names appear once each.

Convergence appeared even before the penultimate link. Going one step further back, to people two removes from the target, we find that the 64 chains passed through 55 individuals. One man, Mr. B, appeared 5 times, and on all occasions sent the document to Mr. G. Other individuals appeared two or three times each.

Additional Characteristics of Chains. Eighty-six per cent of the participants sent the folder to persons they described as friends and acquaintances; 14 per cent sent it to relatives. The same percentages had been observed in an earlier pilot study.

Data on patterns of age, sex and occupation support the plausible hypothesis that participants select recipients from a pool of individuals similar to themselves. The data on age support the hypothesis unequivocally; the data on sex and occupation are complicated by the characteristics of the target and the special requirement of establishing contact with him.

Age was bracketed into ten-year categories and the ages of those who sent the document tabled against the ages of those to whom they sent it. On inspection the table showed a strong tendency to cluster around the diagonal, and a chi-square test showed the association to be significant at better than the .001 level.

Similarly, the sex of each sender was tabled against the sex of the corresponding recipient. Men were ten times more likely to send the document to other men than to women, while women were equally likely to send the folder to males as to females ($p<.001$). These results were affected by the fact that the target was male. In an earlier pilot study using a female target, both men and women were three times as likely to send the document to members of the same sex as to members of the opposite sex. Thus there appear to be three tendencies governing the sex of the recipient: (1) there is a tendency to send the document to someone of one's own sex, but (2) women are more likely to cross sex lines than men, and (3) there

is a tendency to send the document to someone of the same sex as the target person.

The occupations reported by participants were rated on two components—one of social status and one of "industry" affiliation, that is, the subsector of the economy with which the individual would be likely to deal. The coding system was *ad hoc*, designed to fit the occupational titles supplied by participants. Tabling the status and "industry" ratings for all senders of the document against those of respective recipients, we observed a strong tendency for people to select recipients similar to themselves on both measures ($p<.001$ for both tables). However, the strength of the relationship for industry seemed to be largely due to a tendency for the folder to stay within the finance field once it arrived there, obviously because the target was affiliated with that field. Moreover, the participants in the study were a heavily middle-class sample, and the target was himself a member of that class. Thus there was no need for the document to leave middle-class circles in progressing from starters to target.

When separate contingency tables were constructed for complete and incomplete chains, the above results were obtained for both tables. Similarly, when separate tables were constructed for chains originating in the 3 starting populations, the findings held up in all 3 tables. Thus, controlling for completion of chains or for starting population did not affect the finding of demographic homogeneity within chains.

CONCLUSIONS

The contribution of the study lies in the use of acquaintance chains to extend an individual's contacts to a geographically and socially remote target, and in the sheer size of the population from which members of the chains were drawn. The study demonstrated the feasibility of the "small world" technique, and took a step toward demonstrating, defining and measuring inter-connectedness in a large society.

The theoretical machinery needed to deal with social networks is still in its infancy. The empirical technique of this research has two major contributions to make to the development of that theory. First, it sets an upper bound on the minimum number of intermediaries required to link widely separated Americans. Since subjects cannot always foresee the most efficient path to a target, our trace procedure must inevitably produce chains longer than those generated by an accurate theoretical model which takes full account of all paths emanating from an individual. The mean number of intermediaries observed in this study was somewhat greater than five;

additional research (by Korte and Milgram) indicates that this value is quite stable, even when racial crossover is introduced. Both the magnitude and stability of the parameter need to be accounted for. Second, the study has uncovered several phenomena which future models should explain. In particular, the convergence of communication chains through common individuals is an important feature of small world nets, and it should be accounted for theoretically.

There are many additional lines of empirical research that may be examined with the small world method. As suggested earlier, one general paradigm for research is to vary the characteristics of the starting person and the target. Further, one might systematically vary the information provided about the target in order to determine, on the psychological side, what strategies people employ in reaching a distant target, and on the sociological side, what specific variables are critical for establishing contact between people of given characteristics.

REFERENCES

Abelson, R. P.
1967 "Mathematical models in social psychology." Pp. 1–54 in L. Berkowitz (ed.) Advances in Experimental Social Psychology, Vol. III. New York: Academic Press.

Bailey, N. T. J.
1957 The Mathematical Theory of Epidemics. New York: Hafner.

Coleman, J. S., E. Katz and H. Menzel
1966 Medical Innovation: A Diffusion Study. Indianapolis: Bobbs-Merrill.

Fararo, T. J. and M. H. Sunshine
1964 A Study of a Biased Friendship Net. Syracuse: Youth Development Center, Syracuse University.

Foster, C. C., A. Rapoport and C. J. Orwant
1963 "A study of a large sociogram II. Elimination of free parameters." Behavioral Science 8(January):56–65.

Gurevitch, M.
1961 The Social Structure of Acquaintanceship Networks. Unpublished doctoral dissertation, Cambridge: M.I.T.

Korte, C. and S. Milgram
Acquaintance Links Between White and Negro Populations: Application of the Small World Method. Journal of Personality and Social Psychology (in press).

Milgram, S.
1967 "The small world problem." Psychology Today 1(May):61–67.
1969 "Interdisciplinary thinking and the small world problem." Pp. 103–120 in Muzafer Sherif and Carolyn W. Sherif (eds.) Interdisciplinary Relationships in the Social Sciences. Chicago: Aldine Publishing Company.

Pool, I. and M. Kochen
A Non-Mathematical Introduction to a Mathematical Model. Undated mimeo. Cambridge: M.I.T.

Rapoport, A.
1953 "Spread of information through a population with socio-structural bias." Bulletin of Mathematical Biophysics 15(December):523–543.
1963 "Mathematical models of social interaction." Pp. 493–579 in R. D. Luce, R. R. Bush and E. Galanter (eds.) Handbook of Mathematical Psychology, Vol. II. New York: John Wiley and Sons.

Rapoport, A. and W. J. Horvath
1961 "A study of a large sociogram." Behavioral Science 6(October):279–291.

An Experimental Study of Search in Global Social Networks

Peter Sheridan Dodds,[1] Roby Muhamad,[2] Duncan J. Watts[1,2]*

We report on a global social-search experiment in which more than 60,000 e-mail users attempted to reach one of 18 target persons in 13 countries by forwarding messages to acquaintances. We find that successful social search is conducted primarily through intermediate to weak strength ties, does not require highly connected "hubs" to succeed, and, in contrast to unsuccessful social search, disproportionately relies on professional relationships. By accounting for the attrition of message chains, we estimate that social searches can reach their targets in a median of five to seven steps, depending on the separation of source and target, although small variations in chain lengths and participation rates generate large differences in target reachability. We conclude that although global social networks are, in principle, searchable, actual success depends sensitively on individual incentives.

It has become commonplace to assert that any individual in the world can reach any other individual through a short chain of social ties (*1*, *2*). Early experimental work by Travers and Milgram (*3*) suggested that the average length of such chains is roughly six, and recent theoretical (*4*) and empirical (*4–9*) work has generalized the claim to a wide range of nonsocial networks. However, much about this "small world" hypothesis is poorly understood and empirically unsubstantiated. In particular, individuals in real social networks have only limited, local information about the global social network and, therefore, finding short paths represents a nontrivial search effort (*10–12*). Moreover, and contrary to accepted wisdom, experimental evidence for short global chain lengths is extremely limited (*13–15*). For example, Travers and Milgram report 96 message chains (of which 18 were completed) initiated by randomly selected individuals from a city other than the target's (*3*). Almost all other empirical studies of large-scale networks (*4–9*, *16–19*) have focused either on nonsocial networks or on crude proxies of social interaction such as scientific collaboration, and studies specific to e-mail networks have so far been limited to within single institutions (*20*).

We have addressed these issues by conducting a global, Internet-based social search experiment (*21*). Participants registered online (http://smallworld.sociology.columbia.edu) and were randomly allocated one of 18 target persons from 13 countries (table S1). Targets included a professor at an Ivy League university, an archival inspector in Estonia, a technology consultant in India, a policeman in Australia, and a veterinarian in the Norwegian army. Participants were informed that their task was to help relay a message to their allocated target by passing the message to a social acquaintance whom they considered "closer" than themselves to the target. Of the 98,847 individuals who registered, about 25% provided their personal information and initiated message chains. Because subsequent senders were effectively recruited by their own acquaintances, the participation rate after the first step increased to an average of 37%. Including initial and subsequent senders, data were recorded on 61,168 individuals from 166 countries, constituting 24,163 distinct message chains (table S2). More than half of all participants resided in North America and were middle class, professional, college educated, and Christian, reflecting commonly held notions of the Internet-using population (*22*).

In addition to providing his or her chosen contact's name and e-mail address, each sender was also required to describe how he or she had come to know the person, along with the type and strength of the resulting relationship. Table 1 lists the frequencies with which different types of relationships—classified by type, origin, and strength—were invoked by our population of 61,168 active senders. When passing messages, senders typically used friendships in preference to business or family ties; however, almost half of these friendships were formed through either work or school affiliations. Furthermore, successful chains in comparison with incomplete chains disproportionately involved professional ties (33.9 versus 13.2%) rather than friendship and familial relationships (59.8 versus 83.4%) (table S3). Successful chains were also more likely to entail links that originated through work or higher education (65.1 versus 39.6%) (table S4). Men passed messages more frequently to other men (57%), and women to other women (61%), and this tendency to pass to a same-sex contact was strengthened by about 3% if the target was the same gender as the sender and similarly weakened in the opposite case. Individuals in both successful and unsuccessful chains typically used ties to acquaintances they deemed to be "fairly close." However, in successful chains "casual" and "not close" ties were chosen 15.7 and 5.9% more frequently than in unsuccessful chains (table S5), thus adding support, and some resolution, to the longstanding claim that "weak" ties are disproportionately responsible for social connectivity (*23*).

Senders were also asked why they considered their nominated acquaintance a suitable recipient (Table 2). Two reasons—geographical proximity of the acquaintance to the target and similarity of occupation—accounted for at least half of all choices, in general agreement with previous findings (*24*, *25*). Geography clearly dominated the early stages of a chain (when senders were geographically distant) but after the third step was cited less frequently than other characteristics, of which occupation was the most often cited. In contrast with previous claims (*3*, *12*), the presence of highly connected individuals (hubs) appears to have limited relevance to the kind of social search embodied by our experiment (social search with large associated costs/rewards or otherwise modified individual incentives may behave differently). Participants relatively rarely nominated an acquaintance primarily because he or she had many friends (Table 2, "Friends"), and individuals in successful

Table 1. Type, origin, and strength of social ties used to direct messages. Only the top five categories in the first two columns have been listed. The most useful category of social tie is medium-strength friendships that originate in the workplace.

Type of relationship	%	Origin of relationship	%	Strength of relationship	%
Friend	67	Work	25	Extremely close	18
Relatives	10	School/university	22	Very close	23
Co-worker	9	Family/relation	19	Fairly close	33
Sibling	5	Mutual friend	9	Casual	22
Significant other	3	Internet	6	Not close	4

[1]Institute for Social and Economic Research and Policy, Columbia University, 420 West 118th Street, New York, NY 10027, USA. [2]Department of Sociology, Columbia University, 1180 Amsterdam Avenue, New York, NY 10027, USA.

*To whom correspondence should be addressed. E-mail: djw24@columbia.edu

chains were far less likely than those in incomplete chains to send messages to hubs (1.6 versus 8.2%) (table S6). We also find no evidence of message "funneling" (*3*, *9*) through a single acquaintance of the target: At most 5% of messages passed through a single acquaintance of any target, and 95% of all chains were completed through individuals who delivered at most three messages. We conclude that social search appears to be largely an egalitarian exercise, not one whose success depends on a small minority of exceptional individuals.

Although the average participation rate (about 37%) was high relative to those reported in most e-mail–based surveys (*26*), the compounding effects of attrition over multiple links resulted in exponential attenuation of chains as a function of their length and therefore an extremely low chain completion rate (384 of 24,163 chains reached their targets). Chains may have terminated (i) randomly, because of individual apathy or disinclination to participate (*3*, *27*); (ii) preferentially at longer chain lengths, corresponding to the claim that chains get "lost" or are otherwise unable to reach their targets (*13*); or (iii) preferentially at short chain lengths, because, for example, individuals nearer the target are more likely to continue the chain.

Our findings support the random-failure hypothesis for two reasons. First, with the exception of the first step (which is special because senders register rather than receive a message from an acquaintance), the attrition rate remains almost constant for all chain lengths at which we have a sufficiently large N; hence small confidence intervals (Fig. 1A). Second, senders who did not forward their messages after one week were asked why they had not participated. Less than 0.3% of those contacted claimed that they could not think of an appropriate recipient, suggesting that lack of interest or incentive, not difficulty, was the main reason for chain termination.

To estimate the reachability of all targets, we first aggregate the 384 completed chains across targets (Fig. 1B), finding the average chain length to be $<L> = 4.05$. However, this number is misleading because it represents an average only over the completed chains, and shorter chains are more likely to be completed. An "ideal" frequency distribution of chain lengths $n'(L)$ (i.e., the chain lengths that would be observed in the hypothetical limit of zero attrition) may be estimated by accounting for observed attrition as follows: $n'(L) = n(L)/\Pi_{i=0}^{L-1}(1-r_i)$ (Fig. 1C, bars), where $n(L)$ is the observed number of chains completed after L steps (Fig. 1B) and r_L is the maximum-likelihood attrition rate from step L to step $L + 1$ (Fig. 1A, circles). Using the observed values of r_L, we have reconstructed the most likely ideal distribution $n'(L)$ (Fig. 1C, bars) under our assumption of random attrition. Because the tail of the distribution is poorly specified (owing to the small number of observed chains at large, L), we measure its median L_* rather than its mean. We find $L_* = 7$, and this can be thought of as the typical ideal chain length for a hypothetical average individual. By repeating the above procedure for chains that started and ended in the same country ($L_* = 5$) or in different countries ($L_* = 7$), we can disentangle to some extent the different underlying distributions of chains, yielding an estimated range of typical chain lengths $5 \leq L_* \leq 7$, depending on the geographical separation of source and target.

Although the range of L_* and the variation in attrition rates across targets do not appear great, the compounding effects of attrition over the length of a message chain can nevertheless generate large differences in message completion rates. For example, a decrease of 15% in attrition rates, when compounded over the same ideal distribution with $L_* = 6$, can generate an 800% increase in completion rate. The same attrition rates [e.g., $r_0 = 0.75$, $r_L = 0.63$ ($L \geq 1$)], when applied over chains with $L_* = 5$ and 7, respectively, can lead to completion rates that vary by up to a factor of three.

Taken together, this evidence suggests a mixed picture of search in global social networks. On the one hand, all targets may in fact be reachable from random initial senders in only a few steps, with surprisingly little variation across targets in different countries and professions. On the other hand, small differences in either participation rates or the underlying chain lengths can have a dramatic impact on the apparent reachability of different targets. Target 5 (a professor at a prominent U.S. university) stands out in this respect. Because 85% of senders were college educated and more than half were American, participants may have anticipated little difficulty in reaching him, thus accounting for his chains' attrition rate (54%) being much lower than that of any other target (60 to 68%). Target 5 received a notable 44% of all completed chains, yet this result is consistent with his "true" reachability being little different from that of other targets; his allocated senders may simply have been more confident of success.

Our results therefore suggest that if individuals searching for remote targets do not have sufficient incentives to proceed, the small-world hypothesis will not appear to hold (*13*), but that even a slight increase in incentives can render social searches success-

Table 2. Reason for choosing next recipient. All quantities are percentages. Location, recipient is geographically closer; Travel, recipient has traveled to target's region; Family, recipient's family originates from target's region; Work, recipient has occupation similar to target; Education, recipient has similar educational background to target; Friends, recipient has many friends; Cooperative, recipient is considered likely to continue the chain; Other, includes recipient as the target.

L	N	Location	Travel	Family	Work	Education	Friends	Cooperative	Other
1	19,718	33	16	11	16	3	9	9	3
2	7,414	40	11	11	19	4	6	7	2
3	2,834	37	8	10	26	6	6	4	3
4	1,014	33	6	7	31	8	5	5	5
5	349	27	3	6	38	12	6	3	5
6	117	21	3	5	42	15	4	5	5
7	37	16	3	3	46	19	8	5	0

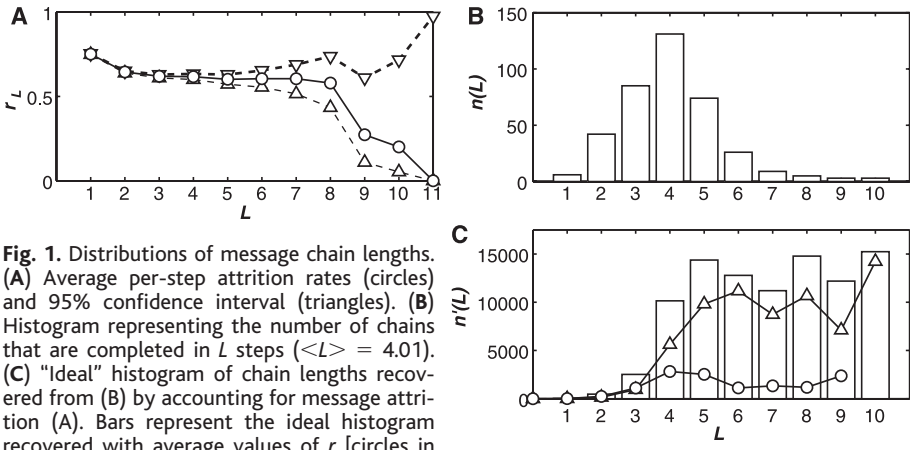


Fig. 1. Distributions of message chain lengths. (**A**) Average per-step attrition rates (circles) and 95% confidence interval (triangles). (**B**) Histogram representing the number of chains that are completed in L steps ($<L> = 4.01$). (**C**) "Ideal" histogram of chain lengths recovered from (B) by accounting for message attrition (A). Bars represent the ideal histogram recovered with average values of r [circles in (A)] for the histogram in (B); lines represent a decomposition of the complete data into chains that start in the same country as the target (circles) and those that start in a different country (triangles).

ful under broad conditions. More generally, the experimental approach adopted here suggests that empirically observed network structure can only be meaningfully interpreted in light of the actions, strategies, and even perceptions of the individuals embedded in the network: Network structure alone is not everything.

References and Notes

1. I. de Sola Pool, M. Kochen, *Soc. Networks* **1**, 1 (1978).
2. S. H. Strogatz, *Nature* **410**, 268 (2001).
3. J. Travers, S. Milgram, *Sociometry* **32**, 425 (1969).
4. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
5. R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130 (1999).
6. L. A. Adamic, in *Lecture Notes in Computer Science 1696*, S. Abiteboul, A. Vercoustre, Eds. (Springer, Heidelberg, 1999), pp. 443–454.
7. L. A. N. Amaral, A. Scala, M. Barthelemy, H. E. Stanley, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11149 (2000).
8. A. Wagner, D. Fell, *Proc. R. Soc. London, B* **268**, 1803 (2001).
9. M. E. J. Newman, *Phys. Rev. E* **64**, 016131 (2001).
10. J. Kleinberg, *Nature* **406**, 845 (2000).
11. D. J. Watts, P. S. Dodds, M. E. J. Newman, *Science* **296**, 1302 (2002).
12. L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman, *Phys. Rev. E* **64**, 046135 (2001).
13. J. S. Kleinfeld, *Society* **39**, 61 (2002).
14. C. Korte, S. Milgram, *J. Pers. Soc. Psychol.* **15**, 101 (1970).
15. N. Lin, P. Dayton, P. Greenwald, in *Communication Yearbook: Vol. 1*, B. D. Ruben, Ed. (Transaction Books, New Brunswick, NJ, 1977), pp. 107–119.
16. A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).
17. M. Faloutsos, P. Faloutsos, C. Faloutsos, *Comp. Comm. Rev.* **29**, 251 (1999).
18. L. A. Adamic, B. A. Huberman, *Science* **287**, 2115a (2000).
19. H. Jeong, B. Tombor, R. Albert, Z. N. Oltavi, A.-L. Barabási, *Nature* **407**, 651 (2000).
20. H. Ebel, L.-I. Mielsch, S. Bornholdt, *Phys. Rev. E* **66**, 035103 (2002).
21. Materials and methods are available as supporting material on *Science* Online.
22. W. Chen, J. Boase, B. Wellman, in *The Internet in Everyday Life*, B. Wellman, C. Haythornthwaite, Eds. (Blackwell, Oxford, 2002), pp. 74–113.
23. M. S. Granovetter, *Am. J. Sociol.* **78**, 1360 (1973).
24. P. D. Killworth, H. R. Bernard, *Soc. Networks* **1**, 159 (1978).
25. H. R. Bernard, P. D. Killworth, M. J. Evans, C. McCarty, G. A. Shelly, *Ethnology* **27**, 155 (1988).
26. K. Sheehan, *J. Comput. Mediated Commun.* **6**(2). Available online at www.ascusc.org/jcmc/vol6/issue2/sheehan.html (2001).
27. H. C. White, *Soc. Forces* **49**(2), 259 (1970).

28. This research was supported in part by the National Science Foundation, Intel Corporation, and Office of Naval Research.

Supporting Online Material

www.sciencemag.org/cgi/content/full/301/5634/827/DC1
Methods
Tables S1 to S6

2 December 2002; accepted 23 May 2003

Phylogenetics and the Cohesion of Bacterial Genomes

Vincent Daubin,[1] Nancy A. Moran,[2] Howard Ochman[1]*

Gene acquisition is an ongoing process in many bacterial genomes, contributing to adaptation and ecological diversification. Lateral gene transfer is considered the primary explanation for discordance among gene phylogenies and as an obstacle to reconstructing the tree of life. We measured the extent of phylogenetic conflict and alien-gene acquisition within quartets of sequenced genomes. Although comparisons of complete gene inventories indicate appreciable gain and loss of genes, orthologs available for phylogenetic reconstruction are consistent with a single tree.

In all but the most reduced bacterial genomes, there is a substantial fraction of genes whose distributions and compositional features indicate that they originated by lateral gene transfer (LGT) (*1*). There is also clear evidence of LGT between distantly related organisms based on phylogenetic studies involving large taxonomic samples (*2*). Given these findings, incompatibility of phylogenies within and among bacterial phyla based on different genes has routinely been ascribed to LGT (*3–10*). However, building molecular phylogenies for distantly related species is often a difficult task, and choice of phylogenetic methods, genes, or taxa can yield different results. For example, there is still no consensus on the monophyly of rodents (*11*, *12*) or the branching order of amniotes (*13*, *14*), and these groups are young compared to bacterial phyla. In addition, distinguishing between orthologous genes (sequences that trace their divergence to the splitting of organismal lineages) and paralogous (duplicated) genes becomes increasingly difficult when considering more distantly related taxa.

The effects of LGT have been extended from the deepest to the shallowest levels of bacterial relationships. Indeed, the similarities in gene sequence and gene content that define widely accepted bacterial taxa have been proposed to reflect boundaries to gene transfer, rather than vertical transmission and common organismal ancestry (*10*). Thus, LGT may overwhelm attempts to reconstruct the relationships among bacterial taxa. The claim that the history of bacteria might be more faithfully depicted as a net than as a tree (*7*) relies upon the postulate that the substantial incidence of acquired DNA within genomes is the basis for findings of phylogenetic incongruence among genes. However, the genes detected as recently transferred are, by and large, different from those used to build species phylogenies. The former are disproportionately A+T-rich, have restricted phylogenetic distributions, and usually encode accessory functions. In contrast, species phylogenies are based on genes with wide taxonomic distributions and having key roles in cellular processes. However, such differences are often ignored when considering the impact of LGT on bacterial relationships. Although the incidence of recently acquired DNA in bacterial genomes is the most direct indication of extensive LGT among species (*1*), the question of whether the incongruence in gene phylogenies is linked to the amount of new DNA in a genome has not been addressed.

To investigate the relation between DNA acquisition and phylogenetic incongruence, we selected quartets of related, sequenced genomes whose phylogenetic relationships, based on small subunit ribosomal RNA (SSU rRNA) sequences, display the branching topology shown in Fig. 1. For each quartet, we inferred both the number of recently acquired and lost genes (based on their phylogenetic distributions) and the proportion of ortholog phylogenies supporting lateral transfers. We applied a conservative method for identifying orthologs by including only those genes having a single significant match per genome, thus minimizing the risks of including hidden paralogs descending from within-genome duplication events. This contrasts with the commonly used "reciprocal best-hit method" (*15*) to infer orthology, which can yield misleading results (*16*), especially when paralogs experience different evolutionary rates. We retained all quartets of species for which >25% of the genes from the smallest genome were recovered as orthologs. We then tested which of the three possible trees was significantly supported for each ortholog family, using the Shimodaira-Hasegawa (SH) (*17*) test implemented in Tree-puzzle 5.1 (*18*) at the 5% level of significance (*19*). This method tests if an alignment significantly supports a tree by estimating the confidence limits of the likelihood estimates of the topologies.

[1]Department of Biochemistry and Molecular Biophysics, [2]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA.

*To whom correspondence should be addressed. E-mail: hochman@email.arizona.edu

Navigation in a small world

It is easier to find short chains between points in some networks than others.

The small-world phenomenon — the principle that most of us are linked by short chains of acquaintances — was first investigated as a question in sociology[1,2] and is a feature of a range of networks arising in nature and technology[3–5]. Experimental study of the phenomenon[1] revealed that it has two fundamental components: first, such short chains are ubiquitous, and second, individuals operating with purely local information are very adept at finding these chains. The first issue has been analysed[2–4], and here I investigate the second by modelling how individuals can find short chains in a large social network.

I have found that the cues needed for discovering short chains emerge in a very simple network model. This model is based on early experiments[1], in which source individuals in Nebraska attempted to transmit a letter to a target in Massachusetts, with the letter being forwarded at each step to someone the holder knew on a first-name basis. The networks underlying the model follow the 'small-world' paradigm[3]: they are rich in structured short-range connections and have a few random long-range connections.

Long-range connections are added to a two-dimensional lattice controlled by a clustering exponent, α, that determines the probability of a connection between two nodes as a function of their lattice distance (Fig. 1a). Decentralized algorithms are studied for transmitting a message: at each step, the holder of the message must pass it across one of its short- or long-range connections; crucially, this current holder does not know the long-range connections of nodes that have not touched the message. The primary figure of merit for such an algorithm is its expected delivery time T, which represents the expected number of steps needed to forward a message between a random source and target in a network generated according to the model. It is crucial to constrain the algorithm to use only local information — with global knowledge of all connections in the network, the shortest chain can be found very simply[6].

A characteristic feature of small-world networks is that their diameter is exponentially smaller than their size, being bounded by a polynomial in log$N$, where N is the number of nodes. In other words, there is always a very short path between any two nodes. This does not imply, however, that a decentralized algorithm will be able to discover such short paths. My central finding is that there is in fact a unique value of the exponent α at which this is possible.

When $\alpha = 2$, so that long-range connections follow an inverse-square distribution, there is a decentralized algorithm that achieves a very rapid delivery time; T is bounded by a function proportional to $(\log N)^2$. The algorithm achieving this bound is a 'greedy' heuristic: each message holder forwards the message across a connection that brings it as close as possible to the target in lattice distance. Moreover, $\alpha = 2$ is the only exponent at which any decentralized algorithm can achieve a delivery time bounded by any polynomial in log$N$: for every other exponent, an asymptotically much larger delivery time is required, regardless of the algorithm employed (Fig. 1b).

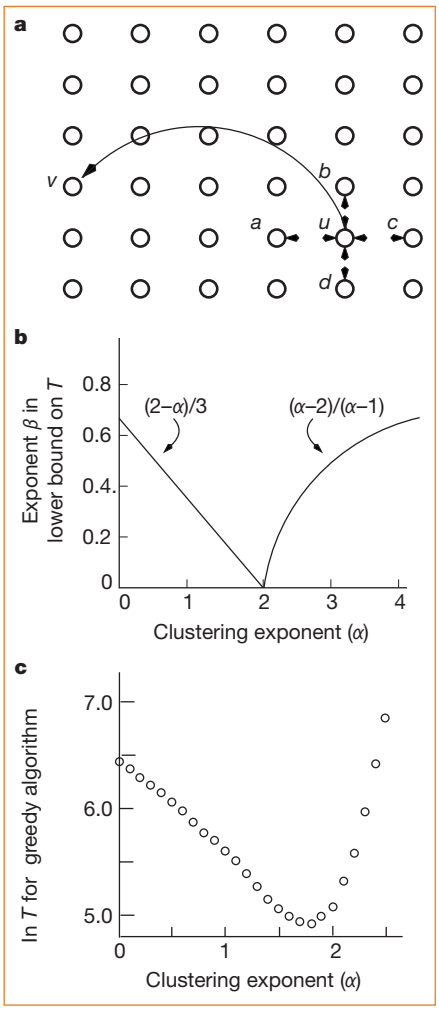


Figure 1 The navigability of small-world networks. **a,** The network model is derived from an $n \times n$ lattice. Each node, u, has a short-range connection to its nearest neighbours (a, b, c and d) and a long-range connection to a randomly chosen node, where node v is selected with probability proportional to $r^{-\alpha}$, where r is the lattice ('Manhattan') distance between u and v, and $\alpha \geq 0$ is a fixed clustering exponent. More generally, for $p,q \geq 1$, each node u has a short-range connection to all nodes within p lattice steps, and q long-range connections generated independently from a distribution with clustering exponent α. **b,** Lower bound from my characterization theorem: when $\alpha \neq 2$, the expected delivery time T of any decentralized algorithm satisfies $T \geq cn^{\beta}$, where $\beta = (2-\alpha)/3$ for $0 \leq \alpha < 2$ and $\beta = (\alpha-2)/(\alpha-1)$ for $\alpha > 2$, and where c depends on α, p and q, but not n. **c,** Simulation of the greedy algorithm on a $20,000 \times 20,000$ toroidal lattice, with random long-range connections as in **a**. Each data point is the average of 1,000 runs.

These results indicate that efficient navigability is a fundamental property of only some small-world structures. The results also generalize to d-dimensional lattices for any value of $d \geq 1$, with the critical value of the clustering exponent becoming $\alpha = d$. Simulations of the greedy algorithm yield results that are qualitatively consistent with the asymptotic analytical bounds (Fig. 1c).

In the areas of communication networks[7] and neuroanatomy[8], the issue of routing without a global network organization has been considered; also in social psychology and information foraging some of the cues that individuals use to construct paths through a social network or hyperlinked environment have been discovered[9,10]. Although I have focused on a very clean model, I believe that a more general conclusion can be drawn for small-world networks — namely that the correlation between local structure and long-range connections provides critical cues for finding paths through the network.

When this correlation is near a critical threshold, the structure of the long-range connections forms a type of gradient that allows individuals to guide a message efficiently towards a target. As the correlation drops below this critical value and the social network becomes more homogeneous, these cues begin to disappear; in the limit, when long-range connections are generated uniformly at random, the result is a world in which short chains exist but individuals, faced with a disorienting array of social contacts, are unable to find them.

Jon M. Kleinberg

Department of Computer Science, Cornell University, Ithaca, New York 14853, USA

1. Milgram, S. *Psychol. Today* **1,** 61–67 (1967).
2. Kochen, M. (ed.) *The Small World* (Ablex, Norwood, NJ, 1989).
3. Watts, D. & Strogatz, S. *Nature* **393,** 440–442 (1998).
4. Albert, R. *et al. Nature* **401,** 130–131 (1999).
5. Adamic, L. in *Proc. 3rd European Conference on Digital Libraries* (eds Abiteboul, S. & Vercoustre, A.-M.) 443–452 (Springer Lecture Notes in Computer Science, Vol. 1696, Berlin, 1999).
6. Cormen, T., Leiserson, C. & Rivest, R. *Introduction to Algorithms* (McGraw-Hill, Boston, 1990).
7. Peleg, D. & Upfal, E. *J. Assoc. Comput. Machinery* **36,** 510–530 (1989).
8. Braitenberg, V. & Schüz, A. *Anatomy of the Cortex* (Springer, Berlin, 1991).
9. Killworth, P. & Bernard, H. *Social Networks* **1,** 159–192 (1978).
10. Pirolli, P. & Card, S. *Psychol. Rev.* **106,** 643–675 (1999).

Identity and search in social networks

Duncan J. Watts,[1, 2, 3, *] Peter Sheridan Dodds,[2, †] and M. E. J. Newman[3, ‡]

[1] *Department of Sociology, Columbia University, New York, NY 10027.*
[2] *Columbia Earth Institute, Columbia University, New York, NY 10027.*
[3] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*

(Dated: January 13, 2006)

Social networks have the surprising property of being "searchable": ordinary people are capable of directing messages through their network of acquaintances to reach a specific but distant target person in only a few steps. We present a model that offers an explanation of social network searchability in terms of recognizable personal identities defined along a number of social dimensions. Our model defines a class of searchable networks and a method for searching them that may be applicable to many network search problems including the location of data files in peer-to-peer networks, pages on the World Wide Web, and information in distributed databases.

In the late 1960's, Travers and Milgram [1] conducted an experiment in which randomly selected individuals in Boston, Massachusetts, and Omaha, Nebraska, were asked to direct letters to a target person in Boston, each forwarding his or her letter to a single acquaintance whom they judged to be closer than themselves to the target. Subsequent recipients did the same. The average length of the resulting acquaintance chains for the letters that eventually reached the target (roughly 20%) was approximately six. This reveals not only that short paths exist [2, 3] between individuals in a large social network but that ordinary people can find these short paths [4]. This is not a trivial statement, since people rarely have more than local knowledge about the network. People know who their friends are. They may also know who some of their friends' friends are. But no one knows the identities of the entire chain of individuals between themselves and an arbitrary target.

The property of being able to find a target quickly, which we call searchability, has been shown to exist in certain specific classes of networks that either possess a certain fraction of hubs (highly connected nodes which, once reached, can distribute messages to all parts of the network [5, 6, 7]) or are built upon an underlying geometric lattice which acts as a proxy for "social space" [4]. Neither of these network types, however, is a satisfactory model of society.

In this paper, we present a model for a social network that is based upon plausible social structures and offers an explanation for the phenomenon of searchability. Our model follows naturally from six contentions about social networks.

1. Individuals in social networks are endowed not only with network ties, but identities [8]: sets of characteristics which they attribute to themselves and others by virtue of their association with, and participation in, social groups [9, 10]. The term group refers to any collection of individuals with which some well-defined set of social characteristics is associated.

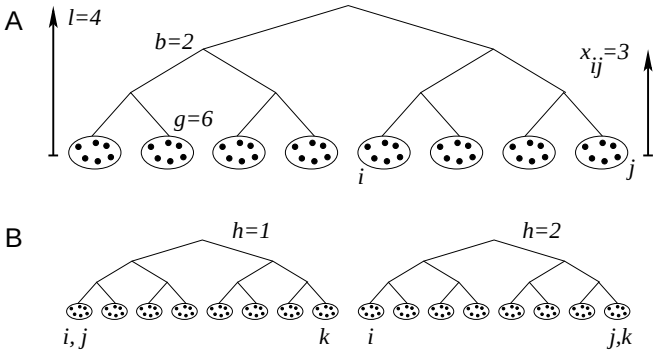


FIG. 1: **(A)** Individuals (dots) belong to groups (ellipses) which in turn belong to groups of groups and so on giving rise to a hierarchical categorization scheme. In this example, groups are composed of $g = 6$ individuals and the hierarchy has $l = 4$ levels with a branching ratio of $b = 2$. Individuals in the same group are considered to be a distance $x = 1$ apart and the maximum separation of two individuals is $x = l$. The example individuals i and j belong to a category two levels above that of their respective groups and the distance between them is $x_{ij} = 3$. Individuals each have z friends in the model and are more likely to be connected with each other the closer their groups are. **(B)** The complete model has many hierarchies indexed by $h = 1 \ldots H$, and the combined social distance y_{ij} between nodes i and j is taken to be the minimum ultrametric distance over all hierarchies $y_{ij} = \min_h x_{ij}^h$. The simple example shown here for $H = 2$ demonstrates that social distance can violate the triangle inequality: $y_{ij} = 1$ since i and j belong to the same group under the first hierarchy and similarly $y_{jk} = 1$ but i and k remain distant in both hierarchies giving $y_{ik} = 4 > y_{ij} + y_{jk} = 2$.

2. Individuals break down, or cluster, the world hierarchically into a series of layers, where the top layer accounts for the entire world and each successively deeper layer represents a cognitive division into a greater number of increasingly specific groups. In principle, this process of distinction by division can be pursued all the way down to the level of individuals, at which point each person is

*Electronic address: djw24@columbia.edu
†Electronic address: p.s.dodds@columbia.edu
‡Electronic address: mark@santafe.edu

uniquely associated with his or her own group. For purposes of identification, however, people do not typically do this, instead terminating the process at the level where the corresponding group size g becomes cognitively manageable. Academic departments, for example, are sometimes small enough to function as a single group, but tend to split into specialized sub-groups as they grow larger. A reasonable upper bound on group size [9] is $g \simeq 100$, a number which we incorporate into our model (Fig. 1A). We define the similarity x_{ij} between individuals i and j as the height of their lowest common ancestor level in the resulting hierarchy, setting $x_{ij} = 1$ if i and j belong to the same group. The hierarchy is fully characterized by depth l and constant branching ratio b. The hierarchy is a purely cognitive construct for measuring social distance and not an actual network. The real network of social connections is constructed as follows.

3. Group membership, in addition to defining individual identity, is a primary basis for social interaction [10, 11], and therefore acquaintanceship. As such, the probability of acquaintance between individuals i and j decreases with decreasing similarity of the groups to which they respectively belong. We model this by choosing an individual i at random and a link distance x with probability $p(x) = c\exp\{-\alpha x\}$, where α is a tunable parameter, and c is a normalizing constant. We then choose a second node j uniformly among all nodes that are distance x from i, repeating this process until we have constructed a network in which individuals have an average number of friends z. The parameter α is therefore a measure of homophily—the tendency of like to associate with like. When $e^{-\alpha} \ll 1$, all links will be as short as possible, and individuals will only connect to those most similar to themselves (i.e., members of their own bottom-level group), yielding a completely homophilous world of isolated cliques. By contrast, when $e^{-\alpha} = b$, any individual is equally likely to interact with any other, yielding a uniform random graph [12] in which the notion of individual similarity or dissimilarity has become irrelevant.

4. Individuals hierarchically cluster the social world in more than one way (for example, by geography and by occupation). We assume that these categories are independent, in the sense that proximity in one does not imply proximity in another. For example, two people may live in the same town but not share the same profession. In our model, we represent each such social dimension by an independently partitioned hierarchy. A node's identity is then defined as an H-dimensional coordinate vector $\vec{v}_i$, where v_i^h is the position of node i in the hth hierarchy, or dimension. Each node i is randomly assigned a coordinate in each of H dimensions, and is then allocated neighbors (friends) as described above, where now it randomly chooses a dimension h (e.g. occupation) to use for each tie. When $H = 1$ and $e^{-\alpha} \ll 1$, the density of network ties must obey the constraint $z < g$.

5. Based on their perceived similarity with other nodes, individuals construct a measure of "social distance" y_{ij}, which we define as the minimum ultrametric distance over all dimensions between two nodes i and j; i.e., $y_{ij} = \min_h x_{ij}^h$. This minimum metric captures the intuitive notion that closeness in only a single dimension is sufficient to connote affiliation (for example, geographically and ethnically distant researchers who collaborate on the same project). A consequence of this minimal metric, depicted in Fig. 1B, is that social distance violates the triangle inequality—hence it is not a true metric distance—because individuals i and j can be close in dimension h_1, and individuals j and k can be close in dimension h_2, yet i and k can be far apart in both dimensions.

6. Individuals forward a message to a single neighbor given only local information about the network. Here, we suppose that each node i knows only its own coordinate vector $\vec{v}_i$, the coordinate vectors $\vec{v}_j$ of its immediate network neighbors, and the coordinate vector of a given target individual $\vec{v}_t$, but is otherwise ignorant of the identities or network ties of nodes beyond its immediate circle of acquaintances.

Individuals therefore have two kinds of partial information: social distance, which can be measured globally but which is not a true distance and hence can yield misleading estimates; and network paths, which generate true distances but which are known only locally. Although neither kind of information alone is sufficient to perform efficient searches, here we show that a simple algorithm that combines knowledge of network ties and social identity can succeed in directing messages with efficiency. The algorithm we implement is the same greedy algorithm Milgram suggested: each member i of a message chain forwards the message to its neighbor j who is perceived to be closer to the target t in terms of social distance; that is, y_{jt} is minimized over all j in i's network neighborhood.

Our principal objective is to determine the conditions under which the average length $\langle L\rangle$ of a message chain connecting a randomly selected sender s to a random target t is small. Although the term small has recently been taken to mean that $\langle L\rangle$ grows slowly with the population size N [13, 14], Travers and Milgram found only that chain lengths were short. Furthermore, these message chains had to be short in an absolute sense because at each step, they were observed to terminate with probability $p \simeq 0.25$ [1, 15]. We therefore adopt a more realistic, functional notion of efficient search, defining for a given message failure probability p, a *searchable network* as any network for which q, the probability of an arbitrary message chain reaching its target, is at least a fixed value r. In terms of chain length, we formally require $q = \langle(1-p)^L\rangle \geq r$, and from this we can obtain an estimate of the maximum required $\langle L\rangle$ using the approximated inequality $\langle L\rangle \leq \ln r/\ln(1-p)$. For the purposes of this paper, we set $r = 0.05$ and $p = 0.25$ giving the stringent requirement that $\langle L\rangle \leq 10.4$ independent of the population size N. Fig. 2A presents a typical phase diagram in H and α outlining the searchable network region

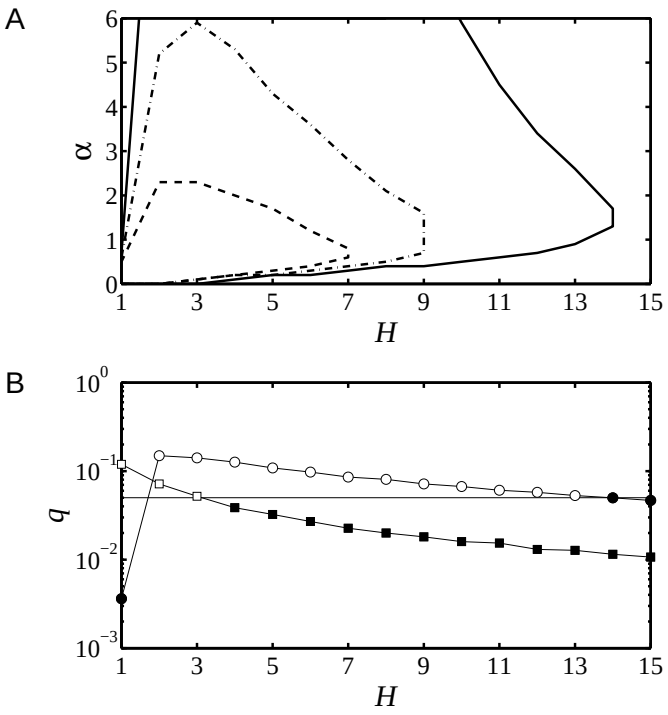


FIG. 2: **(A)** Regions in H-α space where searchable networks exist for varying numbers of individual nodes N (probability of message failure $p = 0.25$, branching ratio $b = 2$, group size $g = 100$, average degree $z = g - 1 = 99$, 10^5 chains sampled per network). The searchability criterion is that the probability of message completion q must be at least $r = 0.05$. The lines correspond to boundaries of the searchable network region for $N = 102400$ (solid), $N = 204800$ (dot-dash), and $N = 409600$ (dash). The region of searchable networks shrinks with N, vanishing at a finite value of N which depends on the model parameters. Note that $z < g$ is required to explore H-α space since for $H = 1$ and α sufficiently large, an individual's neighbors must all be contained within their sole local group. **(B)** Probability of message completion $q(H)$ when $\alpha = 0$ (squares) and $\alpha = 2$ (circles) for the $N = 102400$ data set used in **a**. The horizontal line shows the position of the threshold $r = 0.05$. Open symbols indicate the network is searchable ($q \geq r$) and closed symbols mean otherwise. For $\alpha = 0$, searchability degrades with each additional hierarchy. For the homophilous case of $\alpha = 2$ with a single hierarchy, less than one percent of all searches find their target ($q \simeq 0.004$). Adding just one other hierarchy increases the success rate to $q \simeq 0.144$ and q slowly decreases with H thereafter.

for several choices of N, $g = 100$, and $z = g - 1 = 99$.

Our main result is that searchable networks occupy a broad region of parameter space (α, H) which, as we argue below, corresponds to choices of the model parameters that are the most sociologically plausible. Hence our model suggests that searchability is a generic property of real-world social networks. We support this claim with some further observations, and demonstrate that our model can account for Milgram's experimental findings.

First, we observe that almost all searchable networks display $\alpha > 0$ and $H > 1$, consistent with the notion that individuals are essentially homophilous (that is, they associate preferentially with like individuals), but judge similarity along more than one social dimension. Neither the precise degree to which they are homophilous, nor the exact number of dimensions they choose to use, appear to be important—almost any reasonable choice will do. The best performance, over the largest interval of α, is achieved for $H = 2$ or 3—an interesting result in light of empirical evidence [16] that individuals across different cultures in small-world experiments typically utilize two or three dimensions when forwarding a message.

Second, as Fig. 2B shows, while increasing the number of independent dimensions from $H = 1$ yields a dramatic reduction in delivery time for values of $\alpha > 0$, this improvement is gradually lost as H is increased further. Hence the window of searchable networks in Fig. 2A exhibits an upper boundary in H. Because ties associated with any one dimension are allocated independently with respect to ties in any other dimension, and because for fixed average degree z, larger H necessarily implies fewer ties per dimension, the network ties become less correlated as H increases. In the limit of large H, the network becomes essentially a random graph (regardless of α) and the search algorithm becomes a random walk. Effective decentralized search therefore requires a balance (albeit a highly forgiving one) of categorical flexibility and constraint.

Finally, by introducing parameter choices that are consistent with Milgram's experiment ($N = 10^8$, $p = 0.25$) [1], as well as with subsequent empirical findings ($z = 300$, $H = 2$)[16, 17], we can compare the distribution of chain lengths in our model with those of Travers and Milgram [1] for plausible values of α and b. As Fig. 3 shows, we obtain $\langle L \rangle \simeq 6.7$ for $\alpha = 1$ and $b = 10$, indicating that our model captures the essence of the real small-world problem.

Although sociological in origin, our model is relevant to a broad class of decentralized search problems, such as peer-to-peer networking, in which centralized servers are excluded either by design or by necessity, and where broadcast-type searches (i.e., forwarding messages to all neighbors rather than just one) are ruled out due to congestion constraints [6]. In essence, our model applies to any data structure in which data elements exhibit quantifiable characteristics analogous to our notion of identity, and similarity between two elements—whether people, music files, web pages, or research reports—can be judged along more than one dimension. One of the principal difficulties with designing robust databases [18] is the absence of a unique classification scheme which all users of the database can apply consistently to place and locate files. Two musical songs, for example, can be similar because they belong to the same genre or because they were created in the same year. Our model transforms this difficulty into an asset, allowing all such clas-

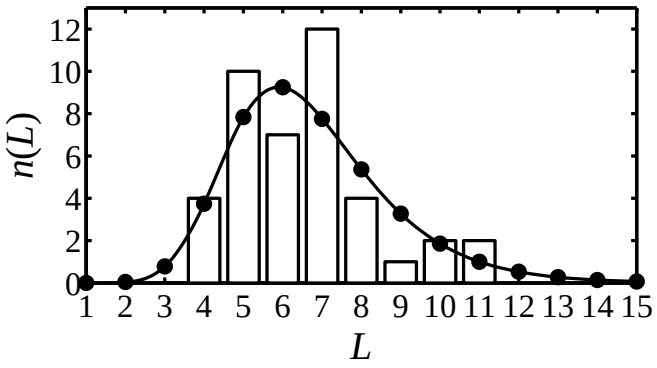


FIG. 3: Comparison between $n(L)$, the number of completed chains of length L, taken from the original small-world experiment [1] (bar graph) and from an example of our model with $N = 10^8$ individuals (filled circles with the line being a guide for the eye). The experimental data shown are for the 42 completed chains that originated in Nebraska. (We have excluded the 24 completed chains that originated in Boston as this would correspond to $N \simeq 10^6$.) The model parameters are $H = 2$, $\alpha = 1$, $b = 10$, $g = 100$, and $z = 300$; message attrition rate is set at 25%; $n(L)$ for the model is compiled from 10^6 random chains and is normalized to match the 42 completed chains that started in Nebraska. The average chain length of Milgram's experiment is approximately 6.5 while the model yields $\langle L \rangle \simeq 6.7$. The distributions compare well: a two-sided Kolmogorov-Smirnov test yields a p-value $P \simeq 0.57$ while for a χ^2 test, $\chi^2 \simeq 5.46$ and $P \simeq 0.49$ (seven bins). (A large value of P supports the hypothesis that the distributions are similar.) Even without attrition, the model's average search time is $\langle L \rangle \simeq 8.5$ and the median chain length is 8. The model does not entirely match the experimental data since the former requires approximately 360 initial chains to achieve 42 completions as compared to 196.

sification schemes to exist simultaneously, and connecting data elements preferentially to similar elements in multiple dimensions. Efficient decentralized searches can then be conducted utilizing simple, greedy algorithms providing only that the characteristics of the target element and the current element's immediate neighbors are known.

Acknowledgments

The authors thank Jon Kleinberg for beneficial discussions. This work was funded in part by the National Science Foundation under grant numbers SES-00-94162 and DMS–0109086, the Intel Corporation, and the Columbia University Office of Strategic Initiatives.

[1] J. Travers and S. Milgram, Sociometry **32**, 425 (1969).
[2] D. J. Watts and S. J. Strogatz, Nature **393**, 440 (1998).
[3] S. H. Strogatz, Nature **410**, 268 (2001).
[4] J. Kleinberg, Nature **406**, 845 (2000).
[5] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[6] L. Adamic, R. Lukose, A. Puniyani, and B. Huberman, Phys. Rev. E **64**, 046135 (2001).
[7] B. J. Kim, C. N. Yoon, S. K. Han, and H. Jeong, Phys. Rev. E (2002), in press.
[8] H. C. White, *Identity and Control* (Princeton University Press, Princeton, 1992).
[9] G. Simmel, American Journal of Sociology **8**, 1 (1902).
[10] F. S. Nadel, *Theory of Social Structure* (Free Press, Glencoe, Ill., 1957).
[11] R. Breiger, Social Forces **53**, 181 (1974).
[12] B. Bollobás, *Random Graphs* (Academic Press, New York, 1985).
[13] M. Newman and D. Watts, Phys. Rev. E **60**, 7332 (1999).
[14] J. Kleinberg, Proc. 32nd ACM Symposium on Theory of Computing (2000).
[15] H. C. White, Social Forces **49**, 259 (1970).
[16] H. Bernard, P. Killworth, M. Evans, C. McCarty, and G. Shelly, Ethnology **27**, 155 (1988).
[17] P. Killworth and H. Bernard, Social Networks **1**, 159 (1978).
[18] B. Manneville, *The Biology of Business: Decoding the Natural Laws of the Enterprise* (Jossey-Bass, San Francisco, 1999), chap. 5. Complex adaptive knowledge management: A case study from McKinsey and Company.

LETTERS

The scaling laws of human travel

D. Brockmann[1,2], L. Hufnagel[3] & T. Geisel[1,2,4]

The dynamic spatial redistribution of individuals is a key driving force of various spatiotemporal phenomena on geographical scales. It can synchronize populations of interacting species, stabilize them, and diversify gene pools[1–3]. Human travel, for example, is responsible for the geographical spread of human infectious disease[4–9]. In the light of increasing international trade, intensified human mobility and the imminent threat of an influenza A epidemic[10], the knowledge of dynamical and statistical properties of human travel is of fundamental importance. Despite its crucial role, a quantitative assessment of these properties on geographical scales remains elusive, and the assumption that humans disperse diffusively still prevails in models. Here we report on a solid and quantitative assessment of human travelling statistics by analysing the circulation of bank notes in the United States. Using a comprehensive data set of over a million individual displacements, we find that dispersal is anomalous in two ways. First, the distribution of travelling distances decays as a power law, indicating that trajectories of bank notes are reminiscent of scale-free random walks known as Lévy flights. Second, the probability of remaining in a small, spatially confined region for a time T is dominated by algebraically long tails that attenuate the super-diffusive spread. We show that human travelling behaviour can be described mathematically on many spatiotemporal scales by a two-parameter continuous-time random walk model to a surprising accuracy, and conclude that human travel on geographical scales is an ambivalent and effectively superdiffusive process.

Quantitative aspects of dispersal in ecology are based on the dispersal curve, which quantifies the relative frequency of travel distances of individuals as a function of geographical distance[11]. A large class of dispersal curves (for example, exponential, gaussian, stretched exponential) permits the identification of a typical length scale by the variance of the displacement length or equivalent quantities. When interpreted as the probability $P(r)$ of finding a displacement of length r in a short time δt, the existence of a typical length scale often justifies the description of dispersal in terms of diffusion equations on large spatiotemporal scales. If, however, $P(r)$ lacks a typical length scale, that is $P(r) \sim r^{-(1+\beta)}$ with $\beta < 2$, the diffusion approximation fails. In physics, random processes with such a single-step distribution are known as Lévy flights[12–16] (see Supplementary Information). Recently, the related notion of long-distance-dispersal (LDD) has been established in dispersal ecology[17], taking into account the observation that dispersal curves of a number of species show power-law tails owing to long-range movements[18–21]. (In ecological literature, the term 'dispersal' is commonly used in the context of the spatial displacement of individuals of a species between their geographical origin of birth and the location of their first breeding place. Here we use the term dispersal to refer to geographical displacements that occur on much shorter timescales, that is, due to travel by various means of transportation.)

Nowadays, humans travel on many spatial scales, ranging from a few to thousands of kilometres over short periods of time. The direct quantitative assessment of human movements, however, is difficult, and a statistically reliable estimate of human dispersal comprising all spatial scales does not exist. The central aim of this work is to use data collected at online bill-tracking websites (which monitor the world-wide dispersal of large numbers of individual bank notes) to infer the statistical properties of human dispersal with very high spatio-temporal precision. Our analysis of human movement is based on the trajectories of 464,670 dollar bills obtained from the bill-tracking system www.wheresgeorge.com. We analysed the dispersal of bank notes in the United States, excluding Alaska and Hawaii. The core data consists of 1,033,095 reports to the bill-tracking website. From these reports we calculated the geographical displacements $r = |\mathbf{x}_2 - \mathbf{x}_1|$ between a first ($\mathbf{x}_1$) and secondary ($\mathbf{x}_2$) report location of a bank note and the elapsed time T between successive reports.

In order to illustrate qualitative features of bank note trajectories, Fig. 1b depicts short-time trajectories ($T < 14$ days) originating from three major US cities: Seattle, New York and Jacksonville. After their initial entry into the tracking system, most bank notes are next reported in the vicinity of the initial entry location, that is $|\mathbf{x}_2 - \mathbf{x}_1| \leq 10$ km (Seattle, 52.7%; New York, 57.7%; Jacksonville, 71.4%). However, a small but considerable fraction is reported beyond a distance of 800 km (Seattle, 7.8%; New York, 7.4%; Jacksonville, 2.9%).

From a total of 20,540 short-time trajectories originating across the United States, we measured the probability $P(r)$ of traversing a distance r in a time interval δT of 1–4 days (Fig. 1c). A total of 14,730 (that is, a fraction $Q = 0.71$) secondary reports occurred outside a short range radius $L_{\min} = 10$ km. Between $L_{\min}$ and the approximate average East–West extension of the United States, $L_{\max} \approx 3{,}200$ km, the kernel shows power-law behaviour $P(r) \sim r^{-(1+\beta)}$ with an exponent $\beta = 0.59 \pm 0.02$. For $r < L_{\min}$, $P(r)$ increases linearly with r, which implies that displacements are distributed uniformly inside the disk $|\mathbf{x}_2 - \mathbf{x}_1| \leq L_{\min}$. We measured $P(r)$ for three classes of initial entry locations: highly populated metropolitan areas (191 sites, local population $N_{\text{loc}} > 120{,}000$), cities of intermediate size (1,544 sites, local population $120{,}000 > N_{\text{loc}} > 22{,}000$) and small towns (23,640 sites, local population $N_{\text{loc}} < 22{,}000$), comprising 35.7%, 29.1% and 25.2% of the entire population of the United States, respectively. The inset in Fig. 1c shows $P(r)$ for these classes. Despite systematic deviations for short distances, all distributions show an algebraic tail with the same exponent $\beta \approx 0.6$, which confirms that the observed power-law is an intrinsic and universal property of dispersal.

However, the situation is more complex. If we assume that the dispersal of bank notes can be described by a Lévy flight with a short-time probability distribution $P(r)$, we can estimate the time T_{eq} for an initially localized ensemble of bank notes to reach the stationary distribution[22] (maps in Fig. 1a), obtaining a value of $T_{\text{eq}} \approx 68$ days (see Supplementary Information). Thus, after 2–3 months, bank notes should have reached an equilibrium distribution. Surprisingly, the long-time dispersal data does not reflect a relaxation within this

[1]Max-Planck Institute for Dynamics and Self-Organisation, Bunsenstr. 10, 37073 Göttingen, Germany. [2]Department of Physics, University of Göttingen, 37073 Göttingen, Germany. [3]Kavli Institute for Theoretical Physics, University of California, Santa Barbara, California 93106, USA. [4]Bernstein Center for Computational Neuroscience, 37073 Göttingen, Germany.

time. Figure 1b shows secondary reports of bank notes with initial entry at Omaha that have dispersed for times $T > 100$ days (with an average time $\langle T \rangle = 289$ days). Only 23.6% of the bank notes travelled farther than 800 km, whereas 57.3% travelled an intermediate distance $50 < r < 800$ km, and a relatively large fraction of 19.1% remained within a radius of 50 km, even after an average time of nearly one year. From the computed value $T_{eq} \approx 68$ days, a much higher fraction of bills is expected to reach the metropolitan areas of the West coast and the New England states after this time. This is sufficient evidence that the simple Lévy flight picture for dispersal is incomplete. What causes this attenuation of dispersal?

Two alternative explanations might account for this effect. The slowing down might be caused by strong spatial inhomogeneities of the system. People might be less likely to leave large cities than for example, suburban areas. Alternatively, long periods of rest might be an intrinsic temporal property of dispersal. In as much as an algebraic tail in spatial displacements yields superdiffusive behaviour, a tail in the probability density $\phi(t)$ for times t between successive spatial displacements of an ordinary random walk can lead to subdiffusion[15] (see Supplementary Information). Here, the ambivalence between scale-free spatial displacements and scale-free periods of rest can be responsible for the observed attenuation of superdiffusion.

In order to address this issue we investigated the relative proportion $P_0^i(t)$ of bank notes which are reported in a small (20 km) radius of the initial entry location i as a function of time (Fig. 1d). The quantity $P_0^i(t)$ estimates the probability of a bank note being reported at the initial location at time t. We computed $P_0^i(t)$ for metropolitan areas, cities of intermediate size and small towns: for all classes we found the asymptotic behaviour $P_0(t) \sim At^{-\eta}$, with the same exponent $\eta = 0.6 \pm 0.03$, which indicates that waiting time and dispersal characteristics are universal. Notice that for a pure Lévy flight with index β in two dimensions, $P_0(t)$ scales with time as $t^{-2/\beta}$ (dashed red line)[15]. For $\beta \approx 0.6$ (as suggested by Fig. 1c) this implies $\eta \approx 3.33$. This is a fivefold steeper decrease than observed, which clearly shows that dispersal cannot be described by a pure Lévy flight. The measured decay is even slower than the decay expected from ordinary two-dimensional diffusion ($\eta = 1$, dashed black line). Therefore, we conclude that the slow decay in $P_0(t)$ reflects the effect

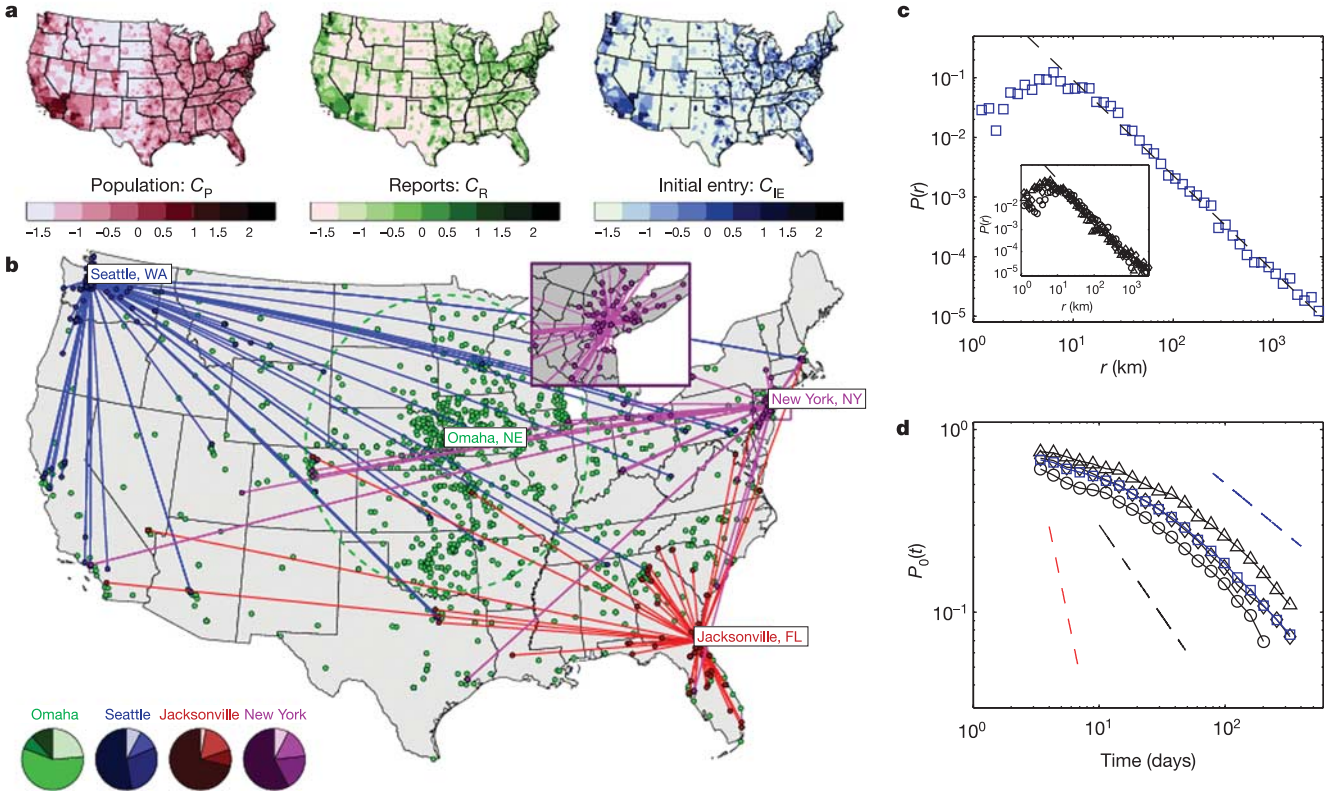


Figure 1 | Dispersal of bank notes and humans on geographical scales. **a**, Relative logarithmic densities of population ($c_P = \log \rho_P / \langle \rho_P \rangle$), report ($c_R = \log \rho_R / \langle \rho_R \rangle$) and initial entry ($c_{IE} = \log \rho_{IE} / \langle \rho_{IE} \rangle$) as functions of geographical coordinates. Colour-code shows densities relative to the nationwide averages (3,109 counties) of $\langle \rho_P \rangle = 95.15$, $\langle \rho_R \rangle = 0.34$ and $\langle \rho_{IE} \rangle = 0.15$ individuals, reports and initial entries per km^2, respectively. **b**, Trajectories of bank notes originating from four different places. City names indicate initial location, symbols secondary report locations. Lines represent short-time trajectories with travelling time $T < 14$ days. Lines are omitted for the long-time trajectories (initial entry in Omaha) with $T > 100$ days. The inset depicts a close-up view of the New York area. Pie charts indicate the relative number of secondary reports coarsely sorted by distance. The fractions of secondary reports that occurred at the initial entry location (dark), at short ($0 < r < 50$ km), intermediate ($50 < r < 800$ km) and long ($r > 800$ km) distances are ordered by increasing brightness of hue. The total number of initial entries are $N = 2{,}055$ (Omaha), $N = 524$ (Seattle), $N = 231$ (New York), $N = 381$ (Jacksonville). **c**, The short-time dispersal kernel. The measured probability density function $P(r)$ of traversing a distance r in less than $T = 4$ days is depicted in blue symbols. It is computed from an ensemble of 20,540 short-time displacements. The dashed black line indicates a power law $P(r) \sim r^{-(1+\beta)}$ with an exponent of $\beta = 0.59$. The inset shows $P(r)$ for three classes of initial entry locations (black triangles for metropolitan areas, diamonds for cities of intermediate size, circles for small towns). Their decay is consistent with the measured exponent $\beta = 0.59$ (dashed line). **d**, The relative proportion $P_0(t)$ of secondary reports within a short radius ($r_0 = 20$ km) of the initial entry location as a function of time. Blue squares show $P_0(t)$ averaged over 25,375 initial entry locations. Black triangles, diamonds, and circles show $P_0(t)$ for the same classes as **c**. All curves decrease asymptotically as $t^{-\eta}$ with an exponent $\eta = 0.60 \pm 0.03$ indicated by the blue dashed line. Ordinary diffusion in two dimensions predicts an exponent $\eta = 1.0$ (black dashed line). Lévy flight dispersal with an exponent $\beta = 0.6$ as suggested by **b** predicts an even steeper decrease, $\eta = 3.33$ (red dashed line).

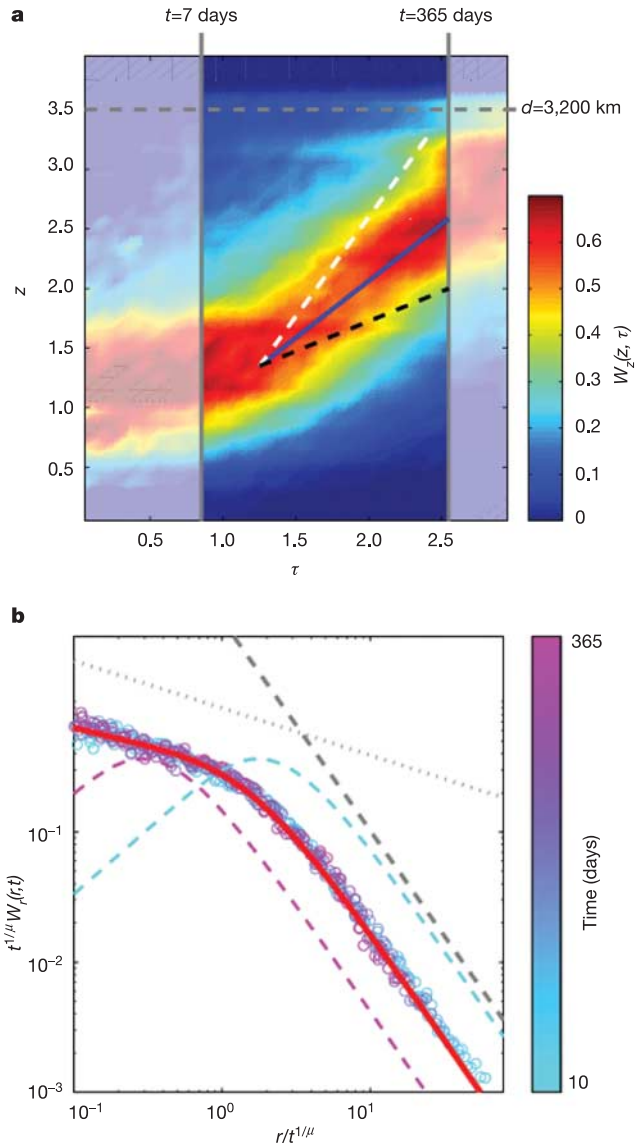


Figure 2 | Spatiotemporal scaling of bank note dispersal. **a**, The probability density $W_z(z,\tau)$ of having travelled a logarithmic distance $z = \log_{10} r$ at logarithmic time $\tau = \log_{10} t$. The middle segment indicates the scaling regime between one week and one year. The superimposed red line represents the scaling behaviour $r(t) \sim t^{1/\mu}$ with exponent $\mu = 1.05 \pm 0.05$. It is compared to the diffusive scaling (black dashed line) and the scaling of a pure Lévy process with exponent $\beta = 0.6$ (white dashed line). The upper dashed grey line shows the approximate linear extent $L_{\max} = 3{,}200$ km of the United States. **b**, The measured radial probability density $W_r(r,t)$ and theoretical scaling function $L_{\alpha,\beta}(r/t^{1/\mu})$ (equation (2)). In order to extract the quality of scaling, the function $t^{1/\mu}W_r(r,t)$ is shown for various but fixed values of t from 10–365 days as a function of the rescaled distance $r/t^{1/\mu}$, where the exponent μ was set to the value determined in **a**. As the measured (circles) curves collapse on a single curve, the process shows universal scaling. The scaling curve represents the limiting density of the process. The asymptotic behaviour for small (grey dotted line) and large (grey dashed line) arguments $y = r/t^{1/\mu}$ is given by $y^{-(1-\xi_1)}$ and $y^{-(1+\xi_2)}$, respectively, with estimated exponents $\xi_1 = 0.63 \pm 0.04$ and $\xi_2 = 0.62 \pm 0.02$. According to our model, these exponents must fulfill $\xi_1 = \xi_2 = \beta$, where β is the exponent of the asymptotic short-time dispersal kernel (Fig. 1c), that is $\beta \approx 0.6$. The superimposed red line represents $t^{1/\mu}W_r(r,t)$ predicted by our theory, with spatial and temporal exponents $\alpha = 0.6$ and $\beta = 0.6$, respectively. The coloured dashed lines represent $t^{1/\mu}W_r(r,t)$ for a pure Lévy flight with $\beta = 0.6$ at times $t = 10$ and $t = 365$ days. The curves do not collapse because the pure Lévy flight shows the wrong spatiotemporal scaling. Furthermore, the limiting curves strongly deviate from the data for small arguments.

of an algebraic tail in the distribution of rests $\phi(t)$ between displacements. Indeed, if $\phi(t) \sim t^{-(1+\alpha)}$ with $\alpha < 1$, then $\eta = \alpha$ and consequently $\alpha = 0.60 \pm 0.03$. This suggests that an algebraic tail in the distribution of rests $\phi(t)$ is responsible for slowing down the superdiffusive dispersal advanced by the short time dispersal kernel in Fig. 1c.

In order to model the antagonistic interplay between scale-free displacements and waiting times, we use the framework of continuous-time random walks (CTRW) introduced by Montroll and Weiss[23]. A CTRW consists of a succession of random displacements $\delta\mathbf{x}_n$ and random waiting times δt_n, each of which is drawn from a corresponding probability density function $P(\delta\mathbf{x}_n)$ and $\phi(\delta t)$. After N iterations, the position of the walker and the elapsed time are given by $\mathbf{x}_N = \sum_n \delta\mathbf{x}_n$ and $t_N = \sum_n \delta t_n$. The quantity of interest is the position $\mathbf{x}(t)$ after time t and the associated probability density $W(\mathbf{x},t)$ that can be computed within CTRW theory. For displacements with finite variance σ^2 and waiting times with finite mean τ, such a CTRW yields ordinary diffusion asymptotically, that is $\partial_t W(\mathbf{x},t) = D\partial_x^2 W(\mathbf{x},t)$ with a diffusion coefficient $D = \sigma^2/\tau$.

In contrast, we assume here that both $P(\delta\mathbf{x}_n)$ and $\phi(\delta t)$ show algebraic tails, that is $P(\delta\mathbf{x}_n) \sim |\delta\mathbf{x}_n|^{-(1+\beta)}$ and $\phi(\delta t) \sim |\delta t|^{-(1+\alpha)}$, for which σ^2 and τ are infinite. In this case we can derive a bifractional diffusion equation for the dynamics of $W(\mathbf{x},t)$:

$$\partial_t^{\alpha} W(\mathbf{x},t) = D_{\alpha,\beta}\partial_{|x|}^{\beta} W(\mathbf{x},t) \qquad (1)$$

In this equation, the symbols ∂_t^{α} and $\partial_{|x|}^{\beta}$ denote fractional derivatives that are non-local and depend on the tail exponents α and β. The constant $D_{\alpha,\beta}$ is a generalized diffusion coefficient (see Supplementary Information). Equation (1) represents the core dynamical equation of our model. Using methods of fractional calculus we can solve this equation and obtain the probability $W_r(r,t)$ of having traversed a distance r at time t:

$$W_r(r,t) = t^{-\alpha/\beta} L_{\alpha,\beta}(r/t^{\alpha/\beta}) \qquad (2)$$

where $L_{\alpha,\beta}$ is a universal scaling function that represents the characteristics of the process. Equation (2) implies that the typical distance travelled scales according to $r(t) \sim t^{1/\mu}$, where $\mu = \beta/\alpha$. Thus, depending on the ratio of spatial and temporal exponents, the random walk can be effectively either superdiffusive ($\beta < 2\alpha$), subdiffusive ($\beta > 2\alpha$), or quasidiffusive ($\beta = 2\alpha$) (see Supplementary Information). For the exponents observed in the dispersal data ($\beta = 0.59 \pm 0.02$ and $\alpha = 0.60 \pm 0.03$) the theory predicts a temporal scaling exponent in the vicinity of unity, $\mu = 0.98 \pm 0.08$. Therefore, dispersal remains superdiffusive despite long periods of rest.

The validity of our model can be tested by estimating $W_r(r,t)$ from the entire data set of a little over half a million displacements and elapse times. The scaling property is best extracted from the data by a transformation to logarithmic coordinates $z = \log_{10} r$, $\tau = \log_{10} t$ and the associated probability density $W_z(z,\tau)$. If the original process scales according to $r(t) \sim t^{1/\mu}$, the density $W_z(z,\tau)$ is a function of $z - \tau/\mu$ only. Figure 2a shows that scaling occurs in a time window of approximately seven days to one year. From the slope (blue line), we obtain a scaling exponent $\mu = 1.05 \pm 0.02$, which agrees well with our model.

Finally, we investigated the degree to which bank note dispersal shows a scaling density as predicted by our model (that is, the relation outlined in equation(2)). Figure 2b shows $t^{1/\mu}W_r(r,t)$ extracted from data versus the ratio $y = r/t^{1/\mu}$. The exponent $\mu = 1.05$ was set to the value obtained in Fig. 2a. The collapse of the data on a single curve indicates that in the chosen time interval of 10–365 days, bank note dispersal shows a universal scaling function. The asymptotic behaviour of the empirical curve is given by $y^{-(1-\xi_1)}$ and $y^{-(1+\xi_2)}$ for small and large arguments, respectively. Both exponents fulfil $\xi_1 \approx \xi_2 \approx 0.6$. We compared the empirical curve with the theoretical prediction of our model. By series expansions, we can compute the asymptotics of the limiting function $L_{\alpha,\beta}(y)$ in equation (2),

giving $y^{-(1-\beta)}$ and $y^{-(1+\beta)}$ for small and large y, respectively. Consequently, as $\beta \approx 0.6$ (Fig. 1c), the theory agrees well will the observed exponents. For the entire range of y we computed $L_{\alpha,\beta}(y)$ by numeric integration for $\alpha = \beta = 0.6$, and superimposed the theoretical curve on the empirical one. The agreement is very satisfactory. In summary, our analysis gives solid evidence that the dispersal of bank notes can be accounted for by our model.

The question remains how the dispersal characteristics of bank notes carry over to the travelling behaviour of humans. In this context, we can conclude that the power law with exponent $\beta = 0.6$ of the short-time dispersal kernel for bank notes reflects the human dispersal kernel, because the exponent remains unchanged for short time intervals of $T = 2$, 4, 7 and 14 days. The issue of long waiting times is more subtle. One might speculate that the observed algebraic tail in waiting times of bank notes is a property of bank note dispersal alone. Long waiting times might be caused by bank notes that exit the money-tracking system for a long time, for instance in banks. However, if this were the case the inter-report time statistics would have an algebraic tail as well. Analysing the inter-report time distribution, we found an exponential decay, suggesting that bank notes are passed from person to person at a constant rate. If we assume that humans exit small areas at a constant rate that is equivalent to exponentially distributed waiting times, and that bank notes pass from person to person at a constant rate, the distribution of bank note waiting times would also be exponential, in contrast to the observed power law. To our minds, this reasoning permits no other conclusion than a lack of scale in human waiting-time statistics. We obtained further support for our results from a comparison with two independent human travel data sets: long-distance travel on the United States aviation network[8] (flight schedules and airport information, www.oag.com; International Air Transport Association, www.iata.org) and the latest survey on long-distance travel conducted by the United States Bureau of Transportation Statistics (www.bts.gov) (see Supplementary Information). Both agree well with our findings and support our conclusions.

On the basis of our analysis, we conclude that the dispersal of bank notes and human travel behaviour can be described by a continuous-time random-walk process that incorporates scale-free jumps as well as long waiting times between displacements. To our knowledge, this is the first empirical evidence for such an ambivalent process in nature. We believe that these results can serve as a starting point for the development of a new class of models for the spread of human infectious diseases, because universal features of human travel can now be accounted for in a quantitative way.

Received 13 July; accepted 3 October 2005.

1. Bullock, J. M., Kenward, R. E. & Hails, R. S. (eds) *Dispersal Ecology* (Blackwell, Malden, Massachusetts, 2002).
2. Murray, J. D. *Mathematical Biology* (Springer-Verlag, New York, 1993).
3. Clobert, J., Danchin, E., Dhondt, A. A. & Nichols, J. D. (eds) *Dispersal* (Oxford Univ. Press, Oxford, 2001).
4. Nicholson, K. & Webster, R. G. *Textbook of Influenza* (Blackwell, Malden, Massachusetts, 1998).
5. Grenfell, B. T., Bjornstadt, O. N. & Kappey, J. Travelling waves and spatial hierarchies in measles epidemics. *Nature* **414,** 716–723 (2001).
6. Keeling, M. J. *et al.* Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* **294,** 813–817 (2001).
7. Hudson, P. J., Rizzoli, A., Grenfell, B. T. & Heesterbeek, H. (eds) *The Ecology of Wildlife Diseases* (Oxford Univ. Press, Oxford, 2002).
8. Hufnagel, L., Brockmann, D. & Geisel, T. Forecast and control of epidemics in a globalized world. *Proc. Natl Acad. Sci. USA* **101,** 15124–15129 (2004).
9. Grassly, N. C., Fraser, C. & Garnett, G. P. Host immunity and synchronized epidemics of syphilis across the United States. *Nature* **433,** 417–421 (2005).
10. Webby, R. J. & Webster, R. G. Are we ready for pandemic influenza? *Science* **302,** 1519–1522 (2003).
11. Kot, M., Lewis, M. A. & van den Driessche, P. Dispersal data and the spread of invading organisms. *Ecology* **77,** 2027–2042 (1996).
12. Shlesinger, M. F., Zaslavsky, G. M. & Frisch, U. (eds) *Lévy Flights and Related Topics in Physics* (Springer Verlag, Berlin, 1995).
13. Klafter, J., Shlesinger, M. F. & Zumofen, G. Beyond Brownian motion. *Phys. Today* **49,** 33–39 (1996).
14. Brockmann, D. & Geisel, T. Lévy flights in inhomogeneous media. *Phys. Rev. Lett.* **90,** 170601 (2003).
15. Metzler, R. & Klafter, J. The random walks guide to anomalous diffusion: a fractional dynamics approach. *Phys. Rep.* **339,** 1–77 (2000).
16. Shlesinger, M. F., Klafter, J. & Wong, Y. M. Random-walks with infinite spatial and temporal moments. *J. Stat. Phys.* **27,** 499–512 (1982).
17. Nathan, R. The challenges of studying dispersal. *Trends Ecol. Evol.* **16,** 481–483 (2001).
18. Viswanathan, G. M. *et al.* Lévy flight search patterns of wandering albatrosses. *Nature* **381,** 413–415 (1996).
19. Ramos-Fernandéz, G., Mateos, J. L., Miramontes, O. & Cocho, G. Lévy walk patterns in the foraging movements of spider monkeys. *Behav. Ecol. Sociobiol.* **55,** 223–230 (2004).
20. Levin, S. A., Muller-Landau, H. C., Nathan, R. & Chave, J. The ecology and evolution of seed dispersal: A theoretical perspective. *Annu. Rev. Ecol. Evol. Syst.* **34,** 575–604 (2003).
21. Nathan, R. *et al.* Mechanisms of long-distance dispersal of seeds by wind. *Nature* **418,** 409–413 (2002).
22. Gardiner, C. W. *Handbook of Stochastic Methods* (Springer Verlag, Berlin, 1985).
23. Montroll, E. W. & Weiss, G. H. Random walks on lattices. *J. Math. Phys.* **6,** 167–181 (1965).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank the initiators of the bill tracking system (www.wheresgeorge.com). We thank cabinetmaker D. Derryberry for discussions and for drawing our attention to the wheresgeorge website, and B. Shraiman, D. Cohen and W. Noyes for critical comments on the manuscript.

Author Contributions The project idea was conceived by D.B. and L.H., data pre-processing was done by L.H., data analysis by D.B. and L.H., the theory and model was constructed by D.B., and the manuscript was written by D.B., L.H. and T.G.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.B. (brockmann@ds.mpg.de).

Four Degrees of Separation

Lars Backstrom* Paolo Boldi† Marco Rosa† Johan Ugander* Sebastiano Vigna†

January 6, 2012

Abstract

Frigyes Karinthy, in his 1929 short story "Láncszemek" ("Chains") suggested that any two persons are distanced by at most six friendship links.[1] Stanley Milgram in his famous experiment [20, 23] challenged people to route postcards to a fixed recipient by passing them only through direct acquaintances. The average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the sample of people chosen.

We report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users (≈ 721 million users, ≈ 69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or "degrees of separation", showing that the world is even smaller than we expected, and prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time.

The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

1 Introduction

At the 20th World–Wide Web Conference, in Hyderabad, India, one of the authors (Sebastiano) presented a new tool for studying the distance distribution of very large graphs: HyperANF [3]. Building on previous graph compression [4] work and on the idea of diffusive computation pioneered in [21], the new tool made it possible to accurately study the distance distribution of graphs orders of magnitude larger than it was previously possible.

One of the goals in studying the distance distribution is the identification of interesting statistical parameters that can be used to tell proper social networks from other complex networks, such as web graphs. More generally, the distance distribution is one interesting *global* feature that makes it possible to reject probabilistic models even when they match local features such as the in-degree distribution.

In particular, earlier work had shown that the *spid*[2], which measures the *dispersion* of the distance distribution, appeared to be smaller than 1 (underdispersion) for social networks, but larger than one (overdispersion) for web graphs [3]. Hence, during the talk, one of the main open questions was "What is the spid of Facebook?".

Lars Backstrom happened to listen to the talk, and suggested a collaboration studying the Facebook graph. This was of course an extremely intriguing possibility: beside testing the "spid hypothesis", computing the distance distribution of the Facebook graph would have been the largest Milgram-like [20] experiment ever performed, orders of magnitudes larger than previous attempts (during our experiments Facebook has ≈ 721 million active users and ≈ 69 billion friendship links).

This paper reports our findings in studying the distance distribution of the largest electronic social network ever created. That world is smaller than we thought: the average distance of the current Facebook graph is 4.74. Moreover, the spid of the graph is just 0.09, corroborating the conjecture [3] that proper social networks have a spid well below one. We also observe, contrary to previous literature analysing graphs orders of magnitude smaller, both a stabilisation of the average distance over time, and that the density of the Facebook graph over time does not neatly fit previous models.

Towards a deeper understanding of the structure of the Facebook graph, we also apply recent compression techniques

*Facebook.

†DSI, Università degli Studi di Milano, Italy. Paolo Boldi, Marco Rosa and Sebastiano Vigna have been partially supported by a Yahoo! faculty grant and by MIUR PRIN "Query log e web crawling".

[1] The exact wording of the story is slightly ambiguous: "He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual [...]". It is not completely clear whether the selected individual is part of the five, so this could actually allude to distance five or six in the language of graph theory, but the "six degrees of separation" phrase stuck after John Guare's 1990 eponymous play. Following Milgram's definition and Guare's interpretation (see further on), we will assume that "degrees of separation" is the same as "distance minus one", where "distance" is the usual path length (the number of arcs in the path).

[2] The spid (shortest-paths index of dispersion) is the variance-to-mean ratio of the distance distribution.

that exploit the underlying cluster structure of the graph to increase *locality*. The results obtained suggests the existence of overlapping clusters similar to those observed in other social networks.

Replicability of scientific results is important. While for obvious nondisclosure reasons we cannot release to the public the actual 30 graphs that have been studied in this paper, we distribute freely the derived data upon which the tables and figures of this papers have been built, that is, the WebGraph *properties*, which contain structural information about the graphs, and the probabilistic estimations of their neighbourhood functions (see below) that have been used to study their distance distributions. The software used in this paper is distributed under the (L)GPL General Public License.[3]

2 Related work

The most obvious precursor of our work is Milgram's celebrated "small world" experiment, described first in [20] and later with more details in [23]: Milgram's works were actually following a stream of research started in sociology and psychology in the late 50s [12]. In his experiment, Milgram aimed at answering the following question (in his words): "given two individuals selected randomly from the population, what is the probability that the minimum number of intermediaries required to link them is 0, 1, 2, ..., k?".

The technique Milgram used (inspired by [22]) was the following: he selected 296 volunteers (the *starting population*) and asked them to dispatch a message to a specific individual (the *target person*), a stockholder living in Sharon, MA, a suburb of Boston, and working in Boston. The message could not be sent directly to the target person (unless the sender knew him personally), but could only be mailed to a personal acquaintance who is more likely than the sender to know the target person. The starting population was selected as follows: 100 of them were people living in Boston, 100 were Nebraska stockholders (i.e., people living far from the target but sharing with him their profession) and 96 were Nebraska inhabitants chosen at random.

In a nutshell, the results obtained from Milgram's experiments were the following: only 64 chains (22%) were completed (i.e., they reached the target); the average number of intermediaries in these chains was 5.2, with a marked difference between the Boston group (4.4) and the rest of the starting population, whereas the difference between the two other subpopulations was not statistically significant; at the other end of the spectrum, the random (and essentially clueless) group from Nebraska needed 5.7 intermediaries on average (i.e., rounding up, "six degrees of separation"). The main conclusions outlined in Milgram's paper were that the average path length is small, much smaller than expected, and that geographic location seems to have an impact on the average length whereas other information (e.g., profession) does not.

There is of course a fundamental difference between our experiment and what Milgram did: Milgram was measuring the average length of a *routing path* on a social network, which is of course an upper bound on the average distance (as the people involved in the experiment were not necessarily sending the postcard to an acquaintance on a shortest path to the destination).[4] In a sense, the results he obtained are even more striking, because not only do they prove that the world is small, but that the actors living in the small world are able to exploit its smallness. It should be remarked, however, that in [20, 23] the purpose of the authors is to estimate the number of intermediaries: the postcards are just a tool, and the details of the paths they follow are studied only as an artifact of the measurement process. The interest in efficient routing lies more in the eye of the beholder (e.g., the computer scientist) than in Milgram's: with at his disposal an actual large database of friendship links and algorithms like the ones we use, he would have dispensed with the postcards altogether.

Incidentally, there have been some attempts to reproduce Milgram-like routing experiments on various large networks [18, 14, 11], but the results in this direction are still very preliminary because notions such as identity, knowledge or routing are still poorly understood in social networks.

We limited ourselves to the part of Milgram's experiment that is more clearly defined, that is, the measurement of shortest paths. The largest experiment similar to the ones presented here that we are aware of is [15], where the authors considered a *communication graph* with 180 million nodes and 1.3 billion edges extracted from a snapshot of the Microsoft Messenger network; they find an average distance of 6.6 (i.e., 5.6 intermediaries; again, rounding up, six degrees of separation). Note, however, that the communication graph in [15] has an edge between two persons only if they communicated during a specific one-month observation period, and thus does not take into account friendship links through which no communication was detected.

The authors of [24], instead, study the distance distribution of some small-sized social networks. In both cases the networks were undirected and small enough (by at least two orders of magnitude) to be accessed efficiently in a random fashion, so the authors used *sampling* techniques. We remark, however, that sampling is not easily applicable to di-

[3] See `http://{webgraph,law}.dsi.unimi.it/`.

[4] Incidentally, this observation is at the basis of one of the most intense monologues in Guare's play: Ouisa, unable to locate Paul, the con man who convinced them he is the son of Sidney Poitier, says "I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. [...] But to find the right six people." Note that this fragment of the monologue clearly shows that Guare's interpretation of the "six degree of separation" idea is equivalent to distance *seven* in the graph-theoretical sense.

rected networks (such as Twitter) that are not strongly connected, whereas our techniques would still work (for some details about the applicability of sampling, see [8]).

Analysing the evolution of social networks in time is also a lively trend of research. Leskovec, Kleinberg and Faloutsos observe in [16] that the average degree of complex networks increase over time while the *effective diameter* shrinks. Their experiments are conducted on a much smaller scale (their largest graph has 4 millions of nodes and 16 millions of arcs), but it is interesting that the phenomena observed seems quite consistent. Probably the most controversial point is the hypothesis that the number of edges $m(t)$ at time t is related to the number of nodes $n(t)$ by the following relation:

$$m(t) \propto n(t)^a,$$

where a is a fixed exponent usually lying in the interval $(1\,..\,2)$. We will discuss this hypothesis in light of our findings.

3 Definitions and Tools

The *neighbourhood function* $N_G(t)$ of a graph G returns for each $t \in \mathbf{N}$ the number of pairs of nodes $\langle x, y\rangle$ such that y is reachable from x in at most t steps. It provides data about how fast the "average ball" around each node expands. From the neighbourhood function it is possible to derive the distance distribution (between reachable pairs), which gives for each t the fraction of reachable pairs at distance exactly t.

In this paper we use HyperANF, a diffusion-based algorithm (building on ANF [21]) that is able to approximate quickly the neighbourhood function of very large graphs; our implementation uses, in turn, WebGraph [4] to represent in a compressed but quickly accessible form the graphs to be analysed.

HyperANF is based on the observation (made in [21]) that $B(x, r)$, the ball of radius r around node x, satisfies

$$B(x, r) = \bigcup_{x \to y} B(y, r-1) \cup \{\, x \,\}.$$

Since $B(x, 0) = \{\, x \,\}$, we can compute each $B(x, r)$ incrementally using sequential scans of the graph (i.e., scans in which we go in turn through the successor list of each node). The obvious problem is that during the scan we need to access randomly the sets $B(x, r-1)$ (the sets $B(x, r)$ can be just saved on disk on a *update file* and reloaded later).

The space needed for such sets would be too large to be kept in main memory. However, HyperANF represents these sets in an *approximate* way, using *HyperLogLog counters* [10], which should be thought as dictionaries that can answer reliably just questions about size. Each such counter is made of a number of small (in our case, 5-bit) *registers*. In a nutshell, a register keeps track of the maximum number M of trailing zeroes of the values of a good hash function applied to the elements of a sequence of nodes: the number of distinct elements in the sequence is then proportional to 2^M. A technique called *stochastic averaging* is used to divide the stream into a number of substreams, each analysed by a different register. The result is then computed by aggregating suitably the estimation from each register (see [10] for details).

The main performance challenge to solve is how to quickly compute the HyperLogLog counter associated to a union of balls, each represented, in turn, by a HyperLogLog counter: HyperANF uses an algorithm based on word-level parallelism that makes the computation very fast, and a carefully engineered implementation exploits multicore architectures with a linear speedup in the number of cores.

Another important feature of HyperANF is that it uses a *systolic* approach to avoid recomputing balls that do not change during an iteration. This approach is fundamental to be able to compute the entire distance distribution, avoiding the arbitrary termination conditions used by previous approaches, which have no provable accuracy (see [3] for an example).

3.1 Theoretical error bounds

The result of a run of HyperANF at the t-th iteration is an estimation of the neighbourhood function in t. We can see it as a random variable

$$\hat{N}_G(t) = \sum_{0 \le i < n} X_{i,t}$$

where each $X_{i,t}$ is the HyperLogLog counter that counts nodes reached by node i in t steps (n is the number of nodes of the graph). When m registers per counter are used, each $X_{i,t}$ has a guaranteed relative standard deviation $\eta_m \le 1.06/\sqrt{m}$.

It is shown in [3] that the output $\hat{N}_G(t)$ of HyperANF at the t-th iteration is an asymptotically almost unbiased estimator of $N_G(t)$, that is

$$\frac{E[\hat{N}_G(t)]}{N_G(t)} = 1 + \delta_1(n) + o(1) \text{ for } n \to \infty,$$

where δ_1 is the same as in [10][Theorem 1] (and $|\delta_1(x)| < 5 \cdot 10^{-5}$ as soon as $m \ge 16$). Moreover, $\hat{N}_G(t)$ has a relative standard deviation not greater than that of the X_i's, that is

$$\frac{\sqrt{\mathrm{Var}[\hat{N}_G(t)]}}{N_G(t)} \le \eta_m.$$

In particular, our runs used $m = 64$ ($\eta_m = 0.1325$) for all graphs except for the two largest Facebook graphs, where we

used $m = 32$ ($\eta_m = 0.187$). Runs were repeated so to obtain a uniform relative standard deviation for all graphs.

Unfortunately, the relative error for the neighbourhood function becomes an *absolute* error for the distance distribution. Thus, the theoretical bounds one obtains for the moments of the distance distribution are quite ugly. Actually, the simple act of dividing the neighbourhood function values by the last value to obtain the cumulative distribution function is nonlinear, and introduces bias in the estimation.

To reduce bias and provide estimates of the standard error of our measurements, we use the *jackknife* [9], a classical nonparametric method for evaluating arbitrary statistics on a data sample, which turns out to be very effective in practice [3].

4 Experiments

The graphs analysed in this paper are graphs of Facebook users who were active in May of 2011; an active user is one who has logged in within the last 28 days. The decision to restrict our study to active users allows us to eliminate accounts that have been abandoned in early stages of creation, and focus on accounts that plausibly represent actual individuals. In accordance with Facebook's data retention policies, historical user activity records are not retained, and historical graphs for each year were constructed by considering currently active users that were registered on January 1st of that year, along with those friendship edges that were formed prior that that date. The "current" graph is simply the graph of active users at the time when the experiments were performed (May 2011). The graph predates the existence of Facebook "subscriptions", a directed relationship feature introduced in August 2011, and also does not include "pages" (such as celebrities) that people may "like". For standard user accounts on Facebook there is a limit of 5 000 possible friends.

We decided to extend our experiments in two directions: regional and temporal. We thus analyse the entire Facebook graph (`fb`), the USA subgraph (`us`), the Italian subgraph (`it`) and the Swedish (`se`) subgraph. We also analysed a combination of the Italian and Swedish graph (`itse`) to check whether combining two regional but distant networks could significantly change the average distance, in the same spirit as in the original Milgram's experiment.[5] For each graph we compute the distance distribution from 2007 up to today by performing several HyperANF runs, obtaining an estimate of values of neighbourhood function with relative standard deviation at most 5.8%: in several cases, however, we performed more runs, obtaining a higher precision. We report the jackknife [9] estimate of derived values (such as average distances) and the associated estimation of the standard error.

4.1 Setup

The computations were performed on a 24-core machine with 72 GiB of memory and 1 TiB of disk space.[6] The first task was to import the Facebook graph(s) into a compressed form for WebGraph [4], so that the multiple scans required by HyperANF's diffusive process could be carried out relatively quickly. This part required some massaging of Facebook's internal IDs into a contiguous numbering: the resulting current `fb` graph (the largest we analysed) was compressed to 345 GB at 20 bits per arc, which is 86% of the information-theoretical lower bound ($\log \binom{n^2}{m}$ bits, there n is the number of nodes and m the number of arcs).[7] Whichever coding we choose, for half of the possible graphs with n nodes and m arcs we need at least $\lfloor \log \binom{n^2}{m} \rfloor$ bits per graph: the purpose of compression is precisely to choose the coding so to represent interesting graphs in a smaller space than that required by the bound.

To understand what is happening, we recall that WebGraph uses the BV compression scheme [4], which applies three intertwined techniques to the successor list of a node:

- successors are (partially) *copied* from previous nodes within a small window, if successors lists are similar enough;
- successors are *intervalised*, that is, represented by a left extreme and a length, if significant contiguous successor sequences appear;
- successors are *gap-compressed* if they pass the previous phases: instead of storing the actual successor list, we store the differences of consecutive successors (in increasing order) using instantaneous codes.

Thus, a graph compresses well when it exhibits *similarity* (nodes with near indices have similar successor lists) and *locality* (successor lists have small gaps).

The better-than-random result above (usually, randomly permuted graphs compressed with WebGraph occupy 10 – 20% more space than the lower bound) has most likely been induced by the renumbering process, as in the original stream of arcs all arcs going out from a node appeared consecutively;

[5] To establish geographic location, we use the users' *current* geo-IP location; this means, for example, that the users in the it-2007 graph are users who are today in Italy and were on Facebook on January 1, 2007 (most probably, American college students then living in Italy).

[6] We remark that the commercial value of such hardware is of the order of a few thousand dollars.

[7] Note that we measure compression with respect to the lower bound on *arcs*, as WebGraph stores *directed* graphs; however, with the additional knowledge that the graph is undirected, the lower bound should be applied to *edges*, thus doubling, in practice, the number of bits used.

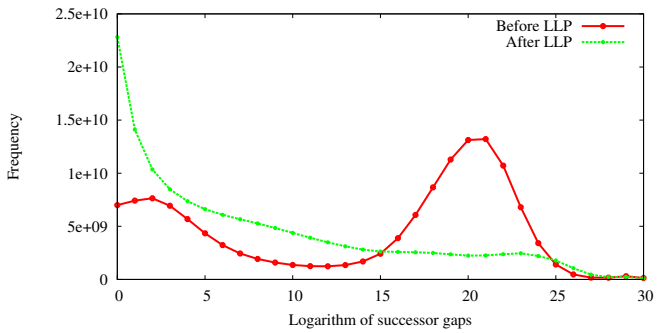


Figure 1: The change in distribution of the logarithm of the gaps between successors when the current `fb` graph is permuted by layered label propagation. See also Table 1.

as a consequence, the renumbering process assigned consecutive labels to all yet-unseen successors (e.g., in the initial stages successors were labelled contiguously), inducing some locality.

It is also possible that the "natural" order for Facebook (essentially, join order) gives rise to some improvement over the information-theoretical lower bound because users often join the network at around the same time as several of their friends, which causes a certain amount of locality and similarity, as circle of friends have several friends in common.

We were interested in the first place to establish whether more locality could be induced by suitably permuting the graph using *layered labelled propagation* [2] (LLP). This approach (which computes several clusterings with different levels of granularity and combines them to sort the nodes of a graph so to increase its locality and similarity) has recently led to the best compression ratios for social networks when combined with the BV compression scheme. An increase in compression means that we were able to partly understand the cluster structure of the graph.

We remark that each of the clusterings required by LLP is in itself a *tour de force*, as the graphs we analyse are almost two orders of magnitude larger than any network used for experiments in the literature on graph clustering. Indeed, applying LLP to the current Facebook graph required ten days of computation on our hardware.

We applied layered labelled propagation and re-compressed our graphs (the current version), obtaining a significant improvement. In Table 1 we show the results: we were able to reduce the graph size by 30%, which suggests that LLP has been able to discover several significant clusters.

The change in structure can be easily seen from Figure 1, where we show the distribution of the binary logarithm of gaps between successors for the current `fb` graph. The smaller the gaps, the higher the locality. In the graph with renumbered Facebook IDs, the distribution is bimodal: there is a local maximum at two, showing that there is some locality, but the bulk of the probability mass is around 20–21, which is slightly less than the information-theoretical lower bound (≈ 23).

In the graph permuted with LLP, however, the distribution radically changes: it is now (mostly) beautifully monotonically decreasing, with a very small bump at 23, which testifies the existence of a small core of "randomness" in the graph that LLP was not able to tame.

Regarding similarity, we see an analogous phenomenon: the number of successors represented by copy has doubled, going from 9% to 18%. The last datum is in line with other social networks (web graphs, on the contrary, are extremely redundant and more than 80% of the successors are usually copied). Moreover, disabling copying altogether results in modest increase in size ($\approx 5\%$), again in line with other social networks, which suggests that for most applications it is better to disable copying at all to obtain faster random access.

The compression ratio is around 53%, which is similar to other similar social networks, such as LiveJournal (55%) or DBLP (40%) [2][8]. For other graphs (see Table 1), however, it is slightly worse. This might be due to several phenomena: First, our LLP runs were executed with only half the number or clusters, and for each cluster we restricted the number of iterations to just four, to make the whole execution of LLP feasible. Thus, our runs are capable of finding considerably less structure than the runs we had previously performed for other networks. Second, the number of nodes is much larger: there is some cost in writing down gaps (e.g., using γ, δ or ζ codes) that is dependent on their absolute magnitude, and the lower bound does not take into account that cost.

4.2 Running

Since most of the graphs, because of their size, had to be accessed by memory mapping, we decided to store all counters (both those for $B(x, r-1)$ and those for $B(x, r)$) in main memory, to avoid eccessive I/O. The runs of HyperANF on the current whole Facebook graph used 32 registers, so the space for counters was about 27 GiB (e.g., we could have analysed a graph with four times the number of nodes on the same hardware). As a rough measure of speed, a run on the LLP-compressed current whole Facebook graph requires about 13.5 hours. Note that this timings would scale linearly with an increase in the number of cores.

4.3 General comments

In September 2006, Facebook was opened to non-college students: there was an instant surge in subscriptions, as our

[8]The interested reader will find similar data for several type of networks at the LAW web site (`http://law.dsi.unimi.it/`).

	it	se	itse	us	fb
Original	14.8 (83%)	14.0 (86%)	15.0 (82%)	17.2 (82%)	20.1 (86%)
LLP	10.3 (58%)	10.2 (63%)	10.3 (56%)	11.6 (56%)	12.3 (53%)

Table 1: The number of bits per link and the compression ratio (with respect to the information-theoretical lower bound) for the current graphs in the original order and for the same graphs permuted by layered label propagation [2].

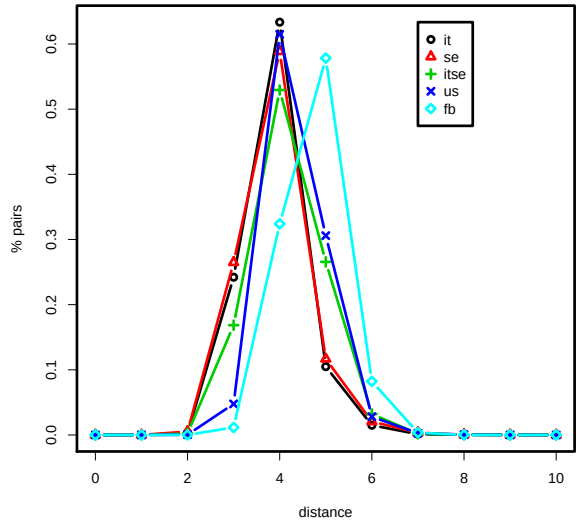


Figure 2: The probability mass functions of the distance distributions of the current graphs (truncated at distance 10).

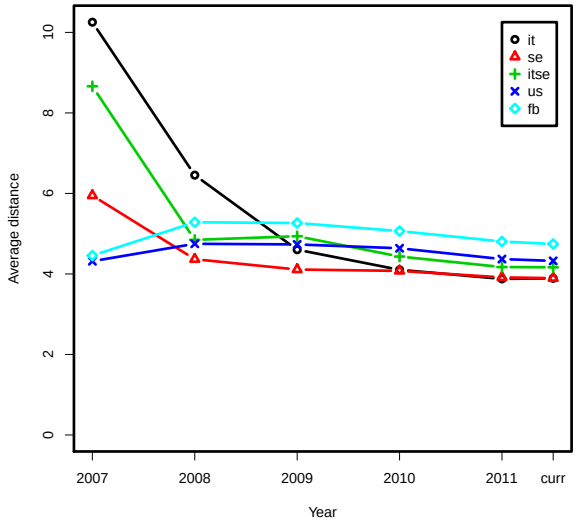


Figure 3: The average distance graph. See also Table 6.

data shows. In particular, the `it` and `se` subgraphs from January 1, 2007 were highly disconnected, as shown by the incredibly low percentage of reachable pairs we estimate in Table 9. Even Facebook itself was rather disconnected, but all the data we compute stabilizes (with small oscillations) after 2009, with essentially all pairs reachable. Thus, we consider the data for 2007 and 2008 useful to observe the evolution of Facebook, but we do not consider them representative of the underlying human social-link structure.

	it	se	itse	us	fb
2007	1.31	3.90	1.50	119.61	99.50
2008	5.88	46.09	36.00	106.05	76.15
2009	50.82	69.60	55.91	111.78	88.68
2010	122.92	100.85	118.54	128.95	113.00
2011	198.20	140.55	187.48	188.30	169.03
current	226.03	154.54	213.30	213.76	190.44

Table 4: Average degree of the datasets.

	it	se	itse	us	fb
2007	0.04	10.23	0.19	100.00	68.02
2008	25.54	93.90	80.21	99.26	89.04

Table 9: Percentage of reachable pairs 2007–2008.

4.4 The distribution

Figure 2 displays the probability mass functions of the current graphs. We will discuss later the variation of the average distance and spid, but qualitatively we can immediately distinguish the *regional* graphs, concentrated around distance four, and the *whole* Facebook graph, concentrated around distance five. The distributions of `it` and `se`, moreover, have significantly less probability mass concentrated on distance five than `itse` and `us`. The variance data (Table 7 and Figure 4) show that the distribution became quickly extremely concentrated.

	it	se	itse	us	fb
2007	159.8 K (105.0 K)	11.2 K (21.8 K)	172.1 K (128.8 K)	8.8 M (529.3 M)	13.0 M (644.6 M)
2008	335.8 K (987.9 K)	1.0 M (23.2 M)	1.4 M (24.3 M)	20.1 M (1.1 G)	56.0 M (2.1 G)
2009	4.6 M (116.0 M)	1.6 M (55.5 M)	6.2 M (172.1 M)	41.5 M (2.3 G)	139.1 M (6.2 G)
2010	11.8 M (726.9 M)	3.0 M (149.9 M)	14.8 M (878.4 M)	92.4 M (6.0 G)	332.3 M (18.8 G)
2011	17.1 M (1.7 G)	4.0 M (278.2 M)	21.1 M (2.0 G)	131.4 M (12.4 G)	562.4 M (47.5 G)
current	19.8 M (2.2 G)	4.3 M (335.7 M)	24.1 M (2.6 G)	149.1 M (15.9 G)	721.1 M (68.7 G)

Table 2: Number of nodes and friendship links of the datasets. Note that each friendship link, being undirected, is represented by a pair of symmetric arcs.

	it	se	itse	us	fb
2007	387.0 K	51.0 K	461.9 K	1.8 G	2.3 G
2008	3.9 M	96.7 M	107.8 M	4.0 G	9.2 G
2009	477.9 M	227.5 M	840.3 M	9.1 G	28.7 G
2010	3.6 G	623.0 M	4.5 G	26.0 G	93.3 G
2011	8.0 G	1.1 G	9.6 G	53.6 G	238.1 G
current	8.3 G	1.2 G	9.7 G	68.5 G	344.9 G

Table 3: Size in bytes of the datasets.

Lower bounds from HyperANF runs

	it	se	itse	us	fb
2007	41	17	41	13	14
2008	28	17	24	17	16
2009	21	16	17	16	15
2010	18	19	19	19	15
2011	17	20	17	18	35
current	19	19	19	20	58
Exact diameter of the giant component					
current	25	23	27	30	41

Table 10: Lower bounds for the diameter of all graphs, and exact values for the giant component ($> 99.7\%$) of current graphs computed using the iFUB algorithm.

4.5 Average degree and density

Table 4 shows the relatively quick growth in time of the average degree of all graphs we consider. The more users join the network, the more existing friendship links are uncovered. In Figure 6 we show a loglog-scaled plot of the same data: with the small set of points at our disposal, it is difficult to draw reliable conclusions, but we are not always observing the power-law behaviour suggested in [16]: see, for instance, the change of the slope for the `us` graph.[9]

The *density* of the network, on the contrary, decreases.[10] In Figure 5 we plot the density (number of edges divided by number of nodes) of the graphs against the number of nodes (see also Table 5). There is some initial alternating behaviour, but on the more complete networks (`fb` and `us`) the trend in sparsification is very evident.

Geographical concentration, however, increases density: in Figure 5 we can see the lines corresponding to our regional graphs clearly ordered by geographical concentration, with the `fb` graph in the lowest position.

4.6 Average distance

The results concerning average distance[11] are displayed in Figure 3 and Table 6. The average distance[12] on the Face-

[9] We remind the reader that on a log-log plot almost anything "looks like" a straight line. The quite illuminating examples shown in [17], in particular, show that goodness-of-fit tests are essential.

[10] We remark that the authors of [16] call *densification* the increase of the average degree, in contrast with established literature in graph theory, where *density* is the fraction of edges with respect to all possible edges (e.g., $2m/(n(n-1))$). We use "density", "densification" and "sparsification" in the standard sense.

[11] The data we report is about the average distance *between reachable pairs*, for which the name *average connected distance* has been proposed [5]. This is the same measure as that used by Travers and Milgram in [23]. We refrain from using the word "connected" as it somehow implies a bidirectional (or, if you prefer, undirected) connection. The notion of average distance between all pairs is useless in a graph in which not all pairs are reachable, as it is necessarily infinite, so no confusion can arise.

[12] In some previous literature (e.g., [16]), the 90% percentile (possibly with some interpolation) of the distance distribution, called *effective diameter*, has been used in place of the average distance. Having at our disposal tools that can compute easily the average distance, which is a parameterless, standard feature of the distance distribution that

	it	se	itse	us	fb
2007	8.224E-06	3.496E-04	8.692E-06	1.352E-05	7.679E-06
2008	1.752E-05	4.586E-05	2.666E-05	5.268E-06	1.359E-06
2009	1.113E-05	4.362E-05	9.079E-06	2.691E-06	6.377E-07
2010	1.039E-05	3.392E-05	7.998E-06	1.395E-06	3.400E-07
2011	1.157E-05	3.551E-05	8.882E-06	1.433E-06	3.006E-07
current	1.143E-05	3.557E-05	8.834E-06	1.434E-06	2.641E-07

Table 5: Density of the datasets.

	it	se	itse	us	fb
2007	10.25 (±0.17)	5.95 (±0.07)	8.66 (±0.14)	4.32 (±0.02)	4.46 (±0.04)
2008	6.45 (±0.03)	4.37 (±0.03)	4.85 (±0.05)	4.75 (±0.02)	5.28 (±0.03)
2009	4.60 (±0.02)	4.11 (±0.01)	4.94 (±0.02)	4.73 (±0.02)	5.26 (±0.03)
2010	4.10 (±0.02)	4.08 (±0.02)	4.43 (±0.03)	4.64 (±0.02)	5.06 (±0.01)
2011	3.88 (±0.01)	3.91 (±0.01)	4.17 (±0.02)	4.37 (±0.01)	4.81 (±0.04)
current	3.89 (±0.02)	3.90 (±0.04)	4.16 (±0.01)	4.32 (±0.01)	4.74 (±0.02)

Table 6: The average distance (± standard error). See also Figure 3 and 7.

book current graph is 4.74.[13] Moreover, a closer look at the distribution shows that 92% of the reachable pairs of individuals are at distance five or less.

We note that both on the `it` and `se` graphs we find a significantly lower, but similar value. We interpret this result as telling us that the average distance is actually dependent on the geographical closeness of users, more than on the actual size of the network. This is confirmed by the higher average distance of the `itse` graph.

During the fastest growing years of Facebook our graphs show a quick decrease in the average distance, which however appears now to be stabilizing. This is not surprising, as "shrinking diameter" phenomena are always observed when a large network is "uncovered", in the sense that we look at larger and larger induced subgraphs of the underlying global human network. At the same time, as we already remarked, density was going down steadily. We thus see the small-world phenomenon fully at work: a smaller fraction of arcs connecting the users, but nonetheless a lower average distance.

To make more concrete the "degree of separation" idea, in Table 11 we show the percentage of reachable pairs *within the ceiling of the average distance* (note, again, that it is the percentage relatively to the reachable pairs): for instance, in the current Facebook graph 92% of the pairs of reachable users are within distance five—four degrees of separation.

has been used in social sciences for decades, we prefer to stick to it. Experimentally, on web and social graphs the average distance is about two thirds of the effective diameter plus one [3].

[13]Note that both Karinthy and Guare had in mind the *maximum*, not the *average* number of degrees, so they were actually upper bounding the diameter.

4.7 Spid

The *spid* is the *index of dispersion* σ^2/μ (a.k.a. *variance-to-mean ratio*) of the distance distribution. Some of the authors proposed the spid [3] as a measure of the "webbiness" of a social network. In particular, networks with a spid larger than one should be considered "web-like", whereas networks with a spid smaller than one should be considered "properly social". We recall that a distribution is called under- or over-dispersed depending on whether its index of dispersion is smaller or larger than 1 (e.g., variance smaller or larger than the average distance), so a network is considered properly social or not depending on whether its distance distribution is under- or over-dispersed.

The intuition behind the spid is that "properly social" networks strongly favour short connections, whereas in the web long connection are not uncommon. As we recalled in the introduction, the starting point of the paper was the question "What is the spid of Facebook"? The answer, confirming the data we gathered on different social networks in [3], is shown in Table 8. With the exception of the highly disconnected regional networks in 2007–2008 (see Table 9), the spid is well below one.

Interestingly, across our collection of graphs we can confirm that there is in general little correlation between the average distance and the spid: Kendall's τ is -0.0105; graphical evidence of this fact can be seen in the scatter plot shown in Figure 7.

If we consider points associated with a single network, though, there appears to be some correlation between average distance and spid, in particular in the more connected

	it	se	itse	us	fb
2007	32.46 (±1.49)	3.90 (±0.12)	16.62 (±0.87)	0.52 (±0.01)	0.65 (±0.02)
2008	3.78 (±0.18)	0.69 (±0.04)	1.74 (±0.15)	0.82 (±0.02)	0.86 (±0.03)
2009	0.64 (±0.04)	0.56 (±0.02)	0.84 (±0.02)	0.62 (±0.02)	0.69 (±0.05)
2010	0.40 (±0.01)	0.50 (±0.02)	0.64 (±0.03)	0.53 (±0.02)	0.52 (±0.01)
2011	0.38 (±0.03)	0.50 (±0.02)	0.61 (±0.02)	0.39 (±0.01)	0.42 (±0.03)
current	0.42 (±0.03)	0.52 (±0.04)	0.57 (±0.01)	0.40 (±0.01)	0.41 (±0.01)

Table 7: The variance of the distance distribution (± standard error). See also Figure 4.

	it	se	itse	us	fb
2007	3.17 (±0.106)	0.66 (±0.016)	1.92 (±0.078)	0.12 (±0.003)	0.15 (±0.004)
2008	0.59 (±0.026)	0.16 (±0.008)	0.36 (±0.028)	0.17 (±0.003)	0.16 (±0.005)
2009	0.14 (±0.007)	0.14 (±0.004)	0.17 (±0.004)	0.13 (±0.003)	0.13 (±0.009)
2010	0.10 (±0.003)	0.12 (±0.005)	0.14 (±0.006)	0.11 (±0.004)	0.10 (±0.002)
2011	0.10 (±0.006)	0.13 (±0.006)	0.15 (±0.004)	0.09 (±0.003)	0.09 (±0.005)
current	0.11 (±0.007)	0.13 (±0.010)	0.14 (±0.003)	0.09 (±0.003)	0.09 (±0.003)

Table 8: The index of dispersion of distances, a.k.a. spid (± standard error). See also Figure 7.

networks (the values for Kendall's τ are all above 0.6, except for `se`). However, this is just an artifact, as the correlation between spid and average distance is *inverse* (larger average distance, smaller spid). What is happening is that in this case the variance (see Table 7) is changing in the same direction: smaller average distances (which would imply a larger spid) are associated with smaller variances. Figure 8 displays the mild correlation between average distance and variance in the graphs we analyse: as a network gets tighter, its distance distribution also gets more concentrated.

4.8 Diameter

HyperANF cannot provide exact results about the diameter: however, the number of steps of a run is necessarily a lower bound for the diameter of the graph (the set of registers can stabilize before a number of iterations equal to the diameter because of hash collisions, but never after). While there are no statistical guarantees on this datum, in Table 10 we report these maximal observations as lower bounds that differ significantly between regional graphs and the overall Facebook graph—there are people that are significantly more "far apart" in the world than in a single nation.[14]

To corroborate this information, we decided to also approach the problem of computing the exact diameter directly, although it is in general a daunting task: for very large graphs matrix-based algorithms are simply not feasible in space, and the basic algorithm running n breadth-first visits is not feasible in time. We thus implemented a highly parallel version of the iFUB (iterative Fringe Upper Bound) algorithm introduced in [6] (extending the ideas of [7, 19]) for undirected graphs.

The basic idea is as follows: consider some node x, and find (by a breadth-first visit) a node y farthest from x. Find now a node z farthest from y: $d(y, z)$ is a (usually very good) lower bound on the diameter, and actually it *is* the diameter if the graph is a tree (this is the "double sweep" algorithm).

We now consider a node c halfway between y and z: such a node is "in the middle of the graph" (actually, it would be a *center* if the graph was a tree), so if h is the eccentricy of c (the distance of the farthest node from c) we expect $2h$ to be a good upper bound for the diameter.

If our upper and lower bound match, we are finished. Otherwise, we consider the *fringe*: the nodes at distance exactly h from c. Clearly, if M is the maximum of the eccentricities of the nodes in the fringe, $\max\{\, 2(h-1), M \,\}$ is a new (and hopefully improved) upper bound, and M is a new (and hopefully improved) lower bound. We then iterate the process by examining fringes closer to the root until the bounds match.

Our implementation uses a multicore breadth-first visit: the queue of nodes at distance d is segmented into small blocks handled by each core. At the end of a round, we have computed the queue of nodes at distance $d + 1$. Our implementation was able to discover the diameter of the current `us` graph (which fits into main memory, thanks to LLP compression) in about twenty minutes. The diameter of Facebook required ten hours of computation of a machine with 1TiB of RAM (actually, 256GiB would have been sufficient, always because of LLP compression).

[14]Incidentally, as we already remarked, this is the measure that Karinthy and Guare actually had in mind.

	it	se	itse	us	fb
2007	65% (11)	64% (6)	67% (9)	95% (5)	91% (5)
2008	77% (7)	93% (5)	77% (5)	83% (5)	91% (6)
2009	90% (5)	96% (5)	75% (5)	86% (5)	94% (6)
2010	98% (5)	97% (5)	91% (5)	91% (5)	97% (6)
2011	90% (4)	86% (4)	95% (5)	97% (5)	89% (5)
current	88% (4)	86% (4)	97% (5)	97% (5)	91% (5)

Table 11: Percentage of reachable pairs within the ceiling of the average distance (shown between parentheses).

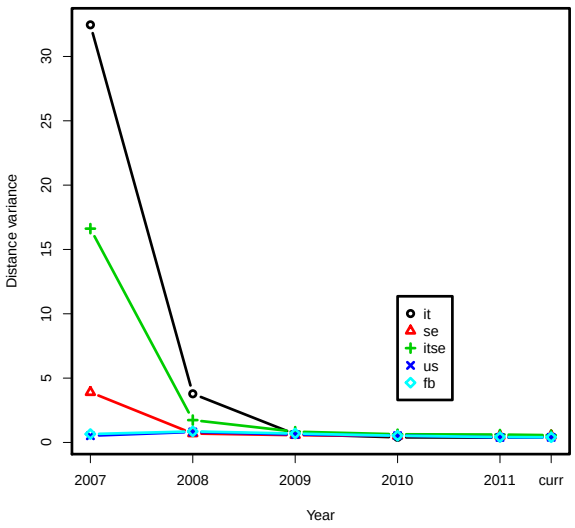


Figure 4: The graph of variances of the distance distributions. See also Table 7.

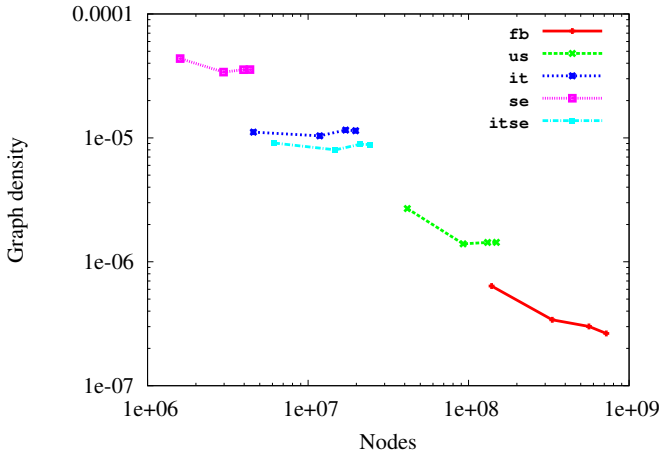


Figure 5: A plot correlating number of nodes to graph density (for the graph from 2009 on).

The values reported in Table 10 confirm what we discovered using the approximate data provided by the length of HyperANF runs, and suggest that while the distribution has a low average distance and it is quite concentrated, there are nonetheless (rare) pairs of nodes that are much farther apart. We remark that in the case of the current `fb` graph, the diameter of the giant component is actually *smaller* than the bound provided by the HyperANF runs, which means that long paths appear in small (and likely very irregular) components.

4.9 Precision

As already discussed in [3], it is very difficult to obtain strong theoretical bounds on data derived from the distance distribution. The problem is that when passing from the neighbourhood function to the distance distribution, the relative error bound becomes an *absolute* error bound: since the dis-

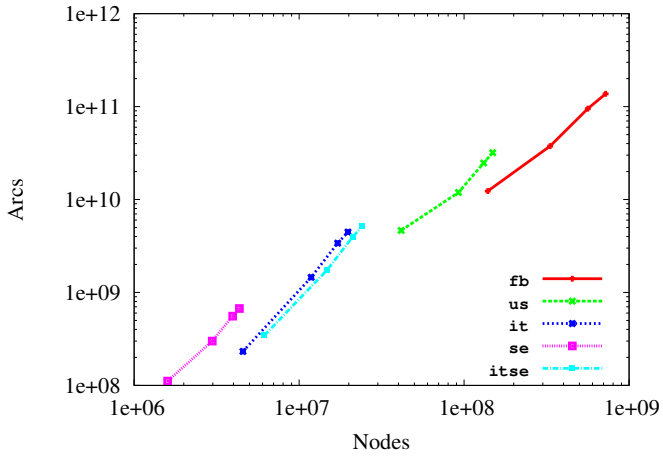


Figure 6: A plot correlating number of nodes to the average degree (for the graphs from 2009 on).

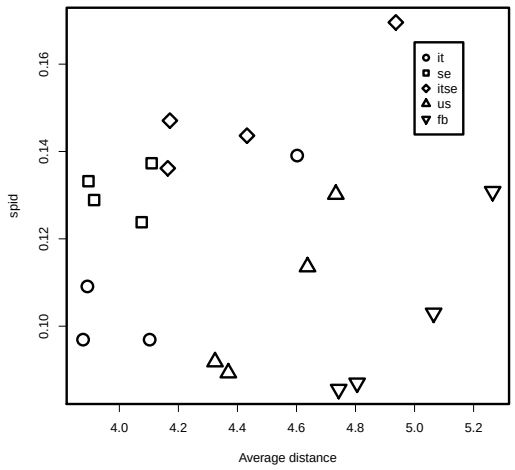


Figure 7: A scatter plot showing the (lack of) correlation between the average distance and the spid.

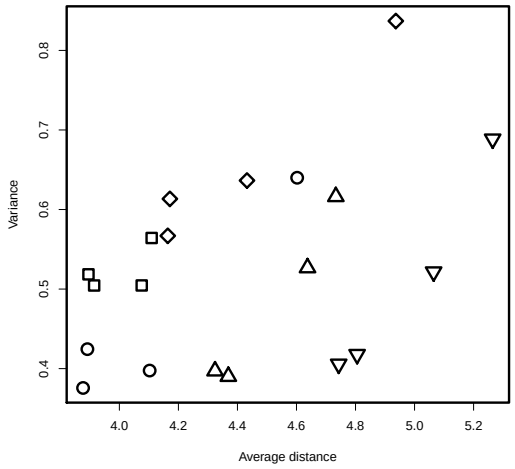


Figure 8: A scatter plot showing the mild correlation between the average distance and the variance.

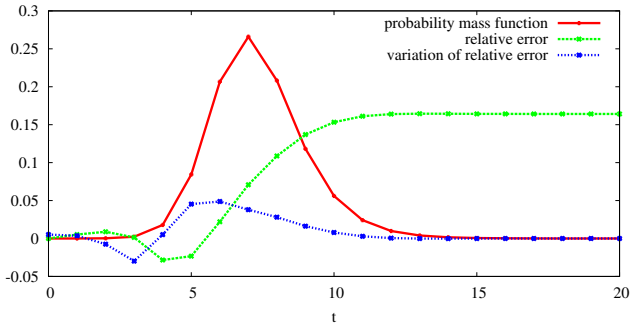


Figure 9: The evolution of the relative error in a HyperANF computation with relative standard deviation 9.25% on a small social network (`dblp-2010`).

tance distribution attains very small values (in particular in its tail), there is a concrete risk of incurring significant errors when computing the average distance or other statistics. On the other hand, the distribution of derived data is extremely concentrated [3].

There is, however, a clear empirical explanation of the unexpected accuracy of our results that is evident from an analysis of the evolution of the empirical relative error of a run on a social network. We show an example in Figure 9.

- In the very first steps, all counters contain essentially disjoint sets; thus, they behave as *independent random variables*, and under this assumption their relative error should be significantly smaller than expected: indeed, this is clearly visible from Figure 9.
- In the following few steps, the distribution reaches its highest value. The error oscillates, as counters are now significantly dependent from one another, but in this part the *actual value of the distribution is rather large*, so the absolute theoretical error turns out to be rather good.
- Finally, in the tail each counter contains a very large subset of the reachable nodes: as a result, all counters behave in a similar manner (as the hash collisions are essentially the same for every counter), and the relative error stabilises to an almost fixed value. Because of this stabilisation, *the relative error on the neighbourhood function transfers, in practice, to a relative error on the distance distribution.* To see why this happen, observe the behaviour of the *variation* of the relative error, which is quite erratic initially, but then converges quickly to zero. The variation is the only part of the relative error that becomes an absolute error when passing to the distance distribution, so the computation on the tail is much more accurate than what the theoretical bound would imply.

We remark that our considerations remain valid for any diffusion-based algorithm using approximate, statistically dependent counters (e.g., ANF [21]).

5 Conclusions

In this paper we have studied the largest electronic social network ever created (≈ 721 million active Facebook users and their ≈ 69 billion friendship links) from several viewpoints.

First of all, we have confirmed that layered labelled propagation [2] is a powerful paradigm for increasing locality of a social network by permuting its nodes. We have been able to compress the `us` graph at 11.6 bits per link—56% of the information-theoretical lower bound, similarly to other, much smaller social networks.

We then analysed using HyperANF the complete Facebook graph and 29 other graphs obtained by restricting geographically or temporally the links involved. We have in fact carried out the largest Milgram-like experiment ever performed. The average distance of Facebook is 4.74, that is, 3.74 "degrees of separation", prompting the title of this paper. The spid of Facebook is 0.09, well below one, as expected for a social network. Geographically restricted networks have a smaller average distance, as it happened in Milgram's original experiment. Overall, these results help paint the picture of what the Facebook social graph looks like. As expected, it is a small-world graph, with short paths between many pairs of nodes. However, the high degree of compressibility and the study of geographically limited subgraphs show that geography plays a huge role in forming the overall structure of network. Indeed, we see in this study, as well as other studies of Facebook [1] that, while the world is connected enough for short paths to exist between most nodes, there is a high degree of locality induced by various externalities, geography chief amongst them, all reminiscent of the model proposed in [13].

When Milgram first published his results, he in fact offered two opposing interpretations of what "six degrees of separation" actually meant. On the one hand, he observed that such a distance is considerably smaller than what one would naturally intuit. But at the same time, Milgram noted that this result could also be interpreted to mean that people are on average six "worlds apart": "When we speak of five[15] intermediaries, we are talking about an enormous psychological distance between the starting and target points, a distance which seems small only because we customarily regard 'five' as a small manageable quantity. We should think of the two points as being not five persons apart, but 'five circles of acquaintances' apart—five 'structures' apart." [20]. From this gloomier perspective, it is reassuring to see that our findings show that people are in fact only four world apart, and not six: when considering another person in the world, a friend of your friend knows a friend of their friend, on average.

[15] Five is the median of the number of intermediaries reported in the first paper by Milgram [20], from which our quotation is taken. More experiments were performed with Travers [23] with a slightly greater average, as reported in Section 2.

References

[1] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[2] Paolo Boldi, Marco Rosa, Massimo Santini, and Sebastiano Vigna. Layered label propagation: A multiresolution coordinate-free ordering for compressing social networks. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proceedings of the 20th international conference on World Wide Web*, pages 587–596. ACM, 2011.

[3] Paolo Boldi, Marco Rosa, and Sebastiano Vigna. HyperANF: Approximating the neighbourhood function of very large graphs on a budget. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proceedings of the 20th international conference on World Wide Web*, pages 625–634. ACM, 2011.

[4] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I: Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference (WWW 2004)*, pages 595–601, Manhattan, USA, 2004. ACM Press.

[5] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web: experiments and models. *Computer Networks*, 33(1–6):309–320, 2000.

[6] P. Crescenzi, R. Grossi, M. Habib, L. Lanzi, and A. Marino. On Computing the Diameter of Real-World Undirected Graphs. Presented at Workshop on Graph Algorithms and Applications (Zurich–July 3, 2011) and selected for submission to the special issue of Theoretical Computer Science in honor of Giorgio Ausiello in the occasion of his 70th birthday, 2011.

[7] Pierluigi Crescenzi, Roberto Grossi, Claudio Imbrenda, Leonardo Lanzi, and Andrea Marino. Finding the diameter in real-world graphs: Experimentally turning a

lower bound into an upper bound. In Mark de Berg and Ulrich Meyer, editors, *Algorithms - ESA 2010, 18th Annual European Symposium, Liverpool, UK, September 6-8, 2010. Proceedings, Part I*, volume 6346 of *Lecture Notes in Computer Science*, pages 302–313. Springer, 2010.

[8] Pierluigi Crescenzi, Roberto Grossi, Leonardo Lanzi, and Andrea Marino. A comparison of three algorithms for approximating the distance distribution in real-world graphs. In Alberto Marchetti-Spaccamela and Michael Segal, editors, *Theory and Practice of Algorithms in (Computer) Systems*, volume 6595 of *Lecture Notes in Computer Science*, pages 92–103. Springer Berlin, 2011.

[9] Bradley Efron and Gail Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

[10] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In *Proceedings of the 13th conference on analysis of algorithm (AofA 07)*, pages 127–146, 2007.

[11] Sharad Goel, Roby Muhamad, and Duncan Watts. Social search in "small-world" experiments. In *Proceedings of the 18th international conference on World wide web*, pages 701–710. ACM, 2009.

[12] Michael Gurevitch. *The social structure of acquaintanceship networks*. PhD thesis, Massachusetts Institute of Technology, Dept. of Economics, 1961.

[13] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[14] Silvio Lattanzi, Alessandro Panconesi, and D. Sivakumar. Milgram-routing in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 725–734. ACM, 2011.

[15] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceeding of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.

[16] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

[17] Lun Li, David L. Alderson, John Doyle, and Walter Willinger. Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Math.*, 2(4), 2005.

[18] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, August 2005.

[19] Clémence Magnien, Matthieu Latapy, and Michel Habib. Fast computation of empirically tight bounds for the diameter of massive graphs. *J. Exp. Algorithmics*, 13:10:1.10–10:1.9, 2009.

[20] Stanley Milgram. The small world problem. *Psychology Today*, 2(1):60–67, 1967.

[21] Christopher R. Palmer, Phillip B. Gibbons, and Christos Faloutsos. Anf: a fast and scalable tool for data mining in massive graphs. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90, New York, NY, USA, 2002. ACM.

[22] Anatol Rapoport and William J. Horvath. A study of a large sociogram. *Behavorial Science*, 6:279–291, October 1961.

[23] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[24] Qi Ye, Bin Wu, and Bai Wang. Distance distribution and average shortest path length estimation in real-world networks. In *Proceedings of the 6th international conference on Advanced data mining and applications: Part I*, volume 6440 of *Lecture Notes in Computer Science*, pages 322–333. Springer, 2010.

DOI:10.1145/2347736.2347753

Human subjects perform a computationally wide range of tasks from only local, networked interactions.

BY MICHAEL KEARNS

Experiments in Social Computation

SINCE 2005, WE have conducted an extensive series of behavioral experiments at the University of Pennsylvania on the ability of human subjects to solve challenging global tasks in social networks from only local, distributed interactions. In these experiments, dozens of subjects simultaneously gather in a laboratory of networked workstations, and are given financial incentives to resolve "their" local piece of some collective problem, which is specified via individual incentives and may involve aspects of coordination, competition, and strategy. The underlying network structures mediating the interaction are unknown to the subjects, and are often chosen from well-studied stochastic models for social network formation. The tasks examined have been drawn from a wide variety of sources, including computer science and complexity theory, game theory and economics, and sociology. They include problems as diverse as graph coloring, networked trading, and biased voting. This article surveys these experiments and their findings.

Our experiments are inherently interdisciplinary, and draw their formulations and motivations from a number of distinct fields. Here, I mention some of these related areas and the questions they have led us to focus upon.

▸ **Computer science.** Within computer science there is current interest in the field's intersection with economics (in the form of algorithmic game theory and mechanism design[22]), including on the topic of strategic interaction in networks, of which our experiments are a behavioral instance. Within the broader technology community, there is also rising interest in the phenomenon of crowdsourcing,[26] citizen science,[18] and related areas, which have yielded impressive "point solutions," but which remains poorly understood in general. What kinds of computational problems can populations of human subjects (perhaps aided by traditional machine resources) solve in a distributed manner from relatively local information and interaction? Does complexity theory or some variant of it provide any guidance? Our experiments have deliberately examined a wide range of problems with varying computational difficulty and strategic properties. In particular, almost all the tasks we have examined entail much more interdependence between user actions than most crowdsourcing efforts to date.

▸ **Behavioral economics and game theory.** Many of our experiments have

» key insights

- **Groups of human subjects are able to solve challenging collective tasks that require considerably more interdependence than most fielded crowdsourcing systems exhibit.**
- **In its current form, computational complexity is a poor predictor of the outcome of our experiments. Equilibrium concepts from economics are more appropriate in some instances.**
- **The possibility of Web-scale versions of our experiments is intriguing, but they will present their own special challenges of subject recruitment, retention, and management.**

ILLUSTRATION BY RANDY LYHUS

an underlying game-theoretic or economic model, and all are conducted via monetary incentives at the level of individual subjects. They can thus be viewed as experiments in behavioral economics,[1] but taking place in (artificial) social networks, an area of growing interest but with little prior experimental literature. In some cases we can make detailed comparisons between behavior and equilibrium predictions, and find systematic (and therefore potentially rectifiable) differences, such as networked instances of phenomena like inequality aversion.

▸ **Network science.** Network Science is itself an interdisciplinary and emerging area[9,25] that seeks to document "universal" structural properties of social and other large-scale networks, and ask how they might form and influence network formation and dynamics. Our experiments can be viewed as extending this line of questioning into a laboratory setting with human subjects, and examining the ways in which network structure influences human behavior, strategies, and performance.

▸ **Computational social science.** While our experimental designs have often emphasized collective problem solving, it is an inescapable fact that individual human subjects make up the collective, and individual decision-making, strategies, and personalities influence the outcomes. What are these influences, and in what ways do they matter? In many of our experiments there are natural and quantifiable notions of traits like stubbornness, stability, and cooperation whose variation across subjects can be measured and correlated with collective behavior and performance, and in turn used to develop simple computational models of individual behavior for predictive and explanatory purposes.

This article surveys our experiments and results to date, emphasizing overall collective performance, behavioral phenomena arising repeatedly across different tasks, task- and network-specific findings that are particularly striking, and the overall methodology and analyses employed. It is worth noting at the outset that one of the greatest challenges posed by this line of work has been the enormous size of the design space: each experimental session involves the selection of a collective problem, a set of network structures, their decomposition into local interactions and subject incentives, and values for many other design variables. Early on we were faced with a choice between breadth and depth—that is, designing experiments to try to populate many points in this space, or picking very specific types of problems and networks, and examining these more deeply over the years. Since the overarching goal of the project has been to explore the broad themes and questions here, and to develop early pieces of a behavioral science of human computation in networked settings, we have opted for breadth, making direct comparisons between some of our experiments difficult. Clearly much more work is needed for a comprehensive picture to emerge.

While our experimental designs have often emphasized collective problem solving, it is an inescapable fact that individual human subjects make up the collective, and individual decision-making, strategies, and personalities influence the outcomes.

In the remainder of this article, I describe the methodology of our experiments, including the system and its GUIs, human subject methodology, and session design. I then summarize our experiments to date and remark on findings that are common to all or most of the different tasks and highlight more specific experimental results on a task-by-task basis.

Experimental Methodology

All of the experiments discussed here were held over a roughly six-year period, in a series of approximately two-hour sessions in the same laboratory of workstations at the University of Pennsylvania. The experiments used an extensive software, network and visualization platform we have developed for this line of research, and which has been used by colleagues at other institutions as well. In all experiments the number of simultaneous subjects was approximately 36, and almost all of the subjects were drawn from Penn undergraduates taking a survey course on the science of social networks.[12] Each experimental session was preceded by a training and demonstration period in which the task, financial incentives, and GUI were explained, and a practice game was held. Sessions were closely proctored to make sure subjects were attending to their workstation and understood the rules and GUI; under no circumstances was advice on strategy provided. Physical partitions were erected around workstations to ensure subjects could only see their own GUI.

No communication or interaction of any kind outside that provided by the system was permitted. The system tabulated the total financial compensation earned by each subject throughout a session, and subjects were paid by check at a later date following the session. Compensation was strictly limited to the actual earnings of each individual subject according to their own play and the rules of the particular task or game; there was no compensation for mere participation. Following a session, subjects were given an exit survey in which they were asked to describe any strategies they employed and behaviors they observed during the experiments.

Within an individual experimental session, the overall collective task or problem was fixed or varied only slightly (for example, an entire session on graph coloring), while the underlying network structures mediating the interaction would vary considerably. Thus, the sessions were structured as a series of short (1 to 5 minutes) experiments, each with its own network structure but on the same task. This is the natural session format, since once the task and incentives are explained to the subjects, it is relatively easy for them to engage in a series of experiments on differing networks, whereas explaining a new task is time-consuming. Each experiment had a time limit imposed by the system, in order to ensure the subjects would not remain stuck indefinitely on any single experiment. In some sessions, there were also conditions for early termination of an experiment, typically when the instance was "solved" (for example, a proper coloring was found). A typical session thus produced between 50 and 100 short experiments.

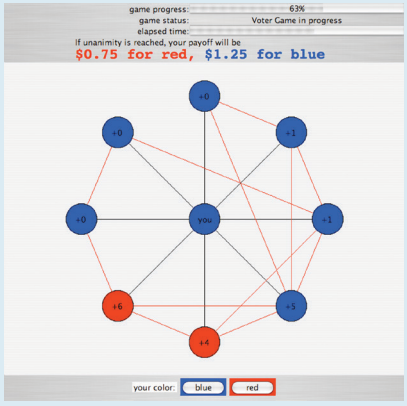
Within an individual experiment, the system randomly assigned subjects to one of the vertices in the network (thus there was neither persistence nor identifiability of network neighbors across experiments). Each subject's GUI (see Figure 1) showed them a local view of the current state of the network—usually a local fragment of the overall network in which the subject's vertex was in the middle and clearly labeled, as well as edges shown to the subject's network neighbors. Edges between a subject's neighbors were shown as well, but no more distant structure. The GUI also always clearly showed the incentives and current payoffs for each subject (which might vary from subject to subject within an experiment), as well the time remaining in the experiment. Typical incentives might pay subjects for being a different color than all their neighbors (graph coloring), the same color (consensus), or perhaps the same color but with different payoffs for different colors (biased voting). Other experiments involved financial scenarios, and the interface provided a mechanism for subjects to bargain or trade with their network neighbors. In general, GUIs always provided enough information for subjects to see the state of their neighbors' current play, and for them to determine their current (financial) best response.

Summary of Experiments

The accompanying table briefly summarizes the nature of the experiments conducted to date, describing the collective task, the network structures used, the individual incentives or mechanism employed, and some of the main findings that we detail below. Our first remark is on the diversity of these experiments along multiple dimensions. In terms of the

Figure 1. Sample screenshot of subject GUI for a biased-voting experiment; many other sessions involved similar GUIs.

The central panel shows the subject's vertex (currently in the "blue" state) with black edges to network neighbors and their current states; red lines denote edges between the subject's neighbors. The bottom action panel allows the subject to change their current state any time, while the top panel specifies their incentives and elapsed time in the experiment.



Summary of experiments to date. ER stands for Erdös-Renyi, PA for preferential attachment.

Task Description	Networks	Incentives/Mechanism	Sample Findings
graph coloring[17]	cycle+chords; PA	differ with neighbors	chords help; importance of information view
coloring and consensus[10]	clique chain w/rewiring	differ/agree with neighbors	opposite structure/task effects
networked trade[13]	ER; PA; structured; all bipartite	limit orders for trades for opposing good	comparison to equilibrium theory; networked inequality aversion
networked bargaining[3]	assorted	Nash bargain on each edge	behavioral price of obstinacy
independent set[15]	assorted	kings and pawns with side payments	side payments help; conflict and fairness
biased voting[14]	ER and PA between types; minority power	consensus with competing individual preferences	well-connected minority rules
network formation[16]	endogenous to the game	biased voting minus edge expenditures	poor collective performance

tasks, the computational complexity of the problems studied[a] varies from the trivial (biased voting and consensus, though this latter problem is difficult in standard models of distributed computation); to the tractable but challenging (networked trade, for which the closest corresponding algorithmic problem is the computation of market equilibria); to the likely intractable (graph coloring and independent set, both *NP*-hard). In terms of the networks, we have investigated standard generative models from the literature such as Erdös-Renyi, preferential attachment, and small worlds; highly structured networks whose design was chosen to highlight strategic tensions in the task and incentives; regular networks without obvious mechanisms to break symmetry; and various other topologies. Figure 2 depicts visualizations of a sampling of network structures investigated. And finally, regarding the financial incentives, these have varied from cooperative (tasks where all players could simultaneously achieve their maximum payoff in the solution); to competitive (where higher payoffs for some players necessarily entail lower payoffs for others); to market-based trading and bargaining, where there are nontrivial networked equilibrium theories and predictions; and to settings where side payments were permitted.

Despite this diversity, and the difficulties in making direct comparisons across sessions and experiments it engenders, there is one unmistakable commonality that has emerged across our six-year investigation: human subjects perform remarkably well at the collective level. While we have observed significant variability in performance across tasks, networks, and incentives, overall the populations have consistently exceeded our expectations. There is a natural and easy way of quantifying this performance: for any given short experiment, we of course know the exact network used, and the incentives and their arrangement within the network, and thus can compute the maximum welfare solution for that particular experiment—that is, the state or arrangement of subject play that would generate the greatest collective payments to the subjects. For each experiment, our system has also recorded the actual payments made, which are by definition less than the maximum social welfare. We can thus sum up all of the actual payments made across all sessions and experiments, and divide it by the sum of all the maximum social welfare payments to arrive at a measure of the overall efficiency of the subject pools over the years.

The resulting figure across the lifetime of our project[b] is 0.88—thus, overall subjects have extracted close to 90% of the payments available to them in principle. In interpreting this figure it should be emphasized that it is an average taken over the particular ensemble of tasks and networks we have studied, which as mentioned before was chosen for its breadth and not in a globally systematic fashion. Clearly it is possible to craft behaviorally "hard" problems and networks.

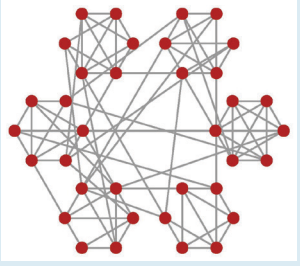
Nevertheless, their efficiency shows that subjects are capable of high performance on a wide variety of tasks and graph topologies.

Another phenomenon consistent across tasks has been the importance of network structure. For most tasks, we found there was a systematic and meaningful dependence of collective behavior on structure, and often an approximate ordering of difficulty of the network topologies could be inferred. Thus, simple cycles prove more difficult for coloring than preferential attachment networks,[17] denser networks result in higher social welfare in networked trading,[13] and so on. However, such dependences on structure are highly task-specific—which is perhaps not surprising for fixed heuristics or algorithms, but has not been documented behaviorally before. Indeed, in one set of experiments we isolated

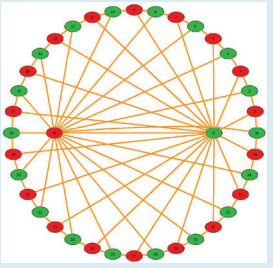
a Clearly computational complexity provides limited insight at best here, since it examines worst-case, centralized, asymptotic computation, all of which are violated in the experiments. But it remains the only comprehensive taxonomy of computational difficulty we have; perhaps these experiments call for a behavioral variant, much as behavioral game theory has provided for its parent field.

b This excludes the most recent experiments in network formation, which are of a qualitatively different nature than the rest, and result in a rather surprising outcome discussed later.

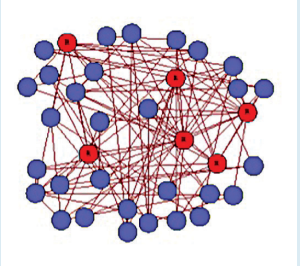
Figure 2. A small sampling of network structures in experiments.



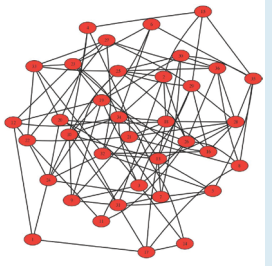
(a) from consensus and independent-set experiments, a chain of six cliques of size 6, with a fraction of the internal clique edges "rewired" to random vertices, thus allowing interpolation between a highly "tribal" network and effectively random networks.



(b) from coloring experiments, an engineered structure with a cycle and two "leaders" in a two-colorable graph.



(c) from biased-voting experiments, a preferential attachment network with a minority of high-degree players preferring red.



(d) from many tasks, a sample Erdös-Renyi network.

this phenomenon by showing that for two cognitively similar (but computationally different) problems, and for a particular generative model for networks, the effects of structure on collective behavioral performance is the opposite in the two tasks,[10] a finding discussed later in greater detail.

The third consistency we found across both tasks and networks was the emergence of individual subject "personalities" or behavioral traits. Our experimental platform is deliberately stylized, and effectively shoehorns the complexity of real human subjects into a highly constrained system, where language, emotion, and other natural forms of communication are eradicated, and all interactions must take place only via simple actions like selecting a color or offering a trade. While there are obvious drawbacks to this stylization in terms of realism, one benefit is that when we make a clear finding—such as the ability of a small but well-connected minority to systematically impose its preferences on the majority[14]—we have done so in a way that might identify the minimal network and task conditions for it to emerge.

Nevertheless, in our experiments we consistently find subjects differentiating and expressing themselves within the constraints of our system in ways that can be measured and compared. For instance, in many of our experiments there are natural notions of traits like stubbornness, stability, selfishness, patience, among others, that can be directly measured in the data, and the frequency of such behavior tallied for each subject. We often find the variation in such behaviors across a population indeed exceeds what can be expected by chance, and thus can be viewed as the personalities of human subjects peeking through our constraints. Harder to measure but still clearly present in almost every experiment we have conducted is the emergence of (sometimes complex) "signaling" mechanisms—it seems that when our system takes language away, the first thing subjects do is try to reintroduce it. From such behavioral traits arise many interesting questions, such as whether specific traits such as stubbornness are correlated with higher payoffs (sometimes they are, other times not), and whether certain mixtures of subject personalities are necessary for effective collective performance (such as a mixture of stubborn and acquiescent individuals in coordination problems).

A theme running throughout our experiments is that intuitions about what networks might be easy or difficult can be strongly violated when considering a distributed human population using only local information.

Highlights of Results

We now turn our attention to results at the level of specific tasks. For each task, I briefly outline any noteworthy details of the GUI or experimental setup, and then highlight some of the main findings.

Coloring and consensus. Our first set of experiments[17] explored the behavioral graph coloring task already alluded to—subjects were given financial incentives to be a different color than their network neighbors, saw only the colors of their local neighborhood, and were free to change their color at any time, choosing from a fixed set of colors whose size was the chromatic number of the underlying graph (thus demanding the subjects find an optimal coloring). It was in these initial experiments that we first found strong effects of network structure. For instance, while a simple two-colorable cycle proved surprisingly hard for the subjects—comparable to their difficulty with more complex and dense preferential attachment graphs—this difficulty was greatly eased by the addition of random chords to the cycle, which reduces diameter and increases edge density. But the preferential attachment networks had the smallest diameter and highest edge density, so these structural properties do not alone explain collective performance.

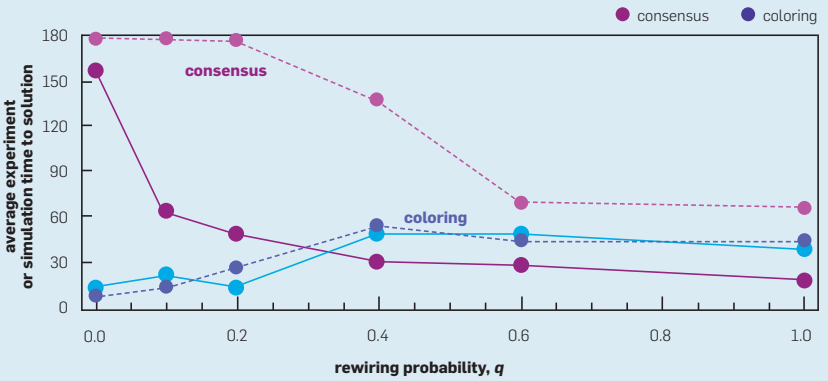
A theme that runs throughout our experiments is that intuitions about what networks might be easy or difficult can be strongly violated when considering a distributed human population using only local information. The challenge of finding simple explanations of such structural results is highlighted by the fact that a natural distributed, randomized heuristic for coloring—namely, not changing colors if there is no current conflict with neighbors, changing to a color resolving a local conflict if one exists, and picking a random color if conflict is unavoidable—produced an ordering of

the difficulty of the networks that was approximately the reverse of that for the subjects.

These first experiments were also the only ones in which we investigated the effects of global information views on performance. In a subset of the experiments, subjects actually saw the current state of the entire network (again with their own vertex in the network clearly indicated), not just the colors of their neighbors. Not surprisingly, this global view led to dramatically improved performance in a simple cycle, where the symmetric structure of the network and the optimal solution become immediately apparent. But strikingly, in preferential attachment networks, global views led to considerable *degradation* in collective performance—perhaps an instance of "information overload," or simply causing subjects to be distracted from attending to their local piece of the global problem.

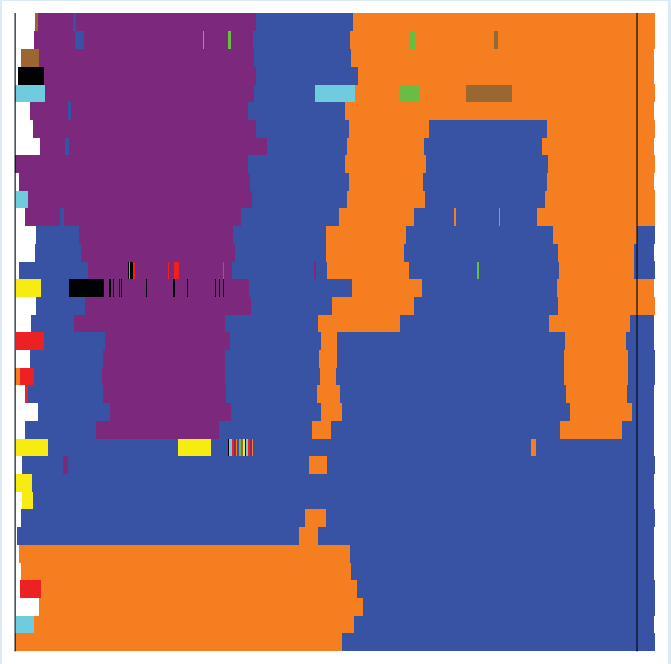
In a later session,[10] we ran experiments on both coloring and *consensus* (where subjects were given financial incentives to be the same color as their neighbors, chosen from a fixed menu of nine colors), on the same set of underlying networks. Despite the vastly different (centralized) computational complexity of these problems—coloring being *NP*-hard, consensus trivial—the two tasks are cognitively very similar and easy for subjects to switch between: coloring is a problem of social differentiation, consensus one of social coordination.

Figure 3. Average time to global solution for coloring and consensus experiments (solid lines) as a function of edge rewiring in a clique-chain network, and simulation times (dashed lines) on the same networks for distributed heuristics. The parametric structure has the opposite effect on the two problems.



In these experiments, the networks were drawn from a parametric family that begins with six cliques of size six loosely connected in a chain. A rewiring parameter q determines the fraction of internal clique edges that are replaced with random "long distance" edges, thus allowing interpolation between a highly clustered, "tribal" network, and the Erdös-Renyi random graph model; see Figure 2(a) for an example. The primary finding here was that the effect on collective performance of varying the rewiring parameter is systematic and *opposite* for the two problems—consensus performance benefits from more rewiring, coloring performance suffers. This effect can be qualitatively captured by simple distributed heuristics, but this does not diminish the striking behavioral phenomenon (see Figure 3). The result suggests that efforts to examine purely structural properties of social and organizational networks, without careful consideration of how structure interacts with the task(s) carried out in those networks, may provide only limited insights on collective behavior.

Figure 4. Visualization of a consensus experiment with low rewiring parameter, showing collective and individual behaviors, and effects of underlying clique structure.



In addition to such systematic, statistically quantifiable results, our experiments often provide interesting opportunities to visualize collective and individual behavior in more anecdotal fashion. Figure 4 shows the actual play during one of the consensus experiments on a network with only a small amount of rewiring, thus largely preserving the tribal clique structure. Each row corresponds to one of the 36 players, and the horizontal axis represents elapsed time in the experiment. The horizontal bars then show the actual color choice by the player at that moment. The first six rows correspond to the players in the first (partially rewired) clique, the next six to the second clique, and so on. The underlying

network structure manifests itself visually in the tendency for these groups of six to change colors approximately simultaneously. As was typical, after an initial diversity of colors, the population quickly settles down to just two or three, and nearly converges to blue before a trickle of orange propagates through the network and takes firm hold; at some point the majority is orange, but this wanes again until the experiment ends in deadlock. Acts of individual signaling (such as toggling between colors) and (apparent) irrationality or experimentation (playing a color not present anywhere else in the network) can also be observed.

Networked trading and bargaining. Our experiments on trading and bargaining differ from the others in that they are accompanied by nontrivial equilibrium theories that generalize certain classical microeconomic models to the networked setting.[4,11] In the networked trading experiments,[4] there were two virtual goods available for trade—call them milk and wheat—and two types of players: those that start with an endowment of milk, but whose payoff is proportional only to how much wheat they obtain via trade; and those that start with wheat but only value milk. All networks were bipartite between the two types of players, and trade was permitted only with network neighbors; players endowed with milk could only trade for wheat and vice-versa, so there were no "resale" or arbitrage opportunities. All endowments were fully divisible and equal, so the only asymmetries are due to network position. The system GUI allowed players to broadcast to their neighbors a proposed rate of exchange[c] of their endowment good for the other good in the form of a traditional limit order in financial markets, and to see the counter offers made by their neighbors; any time the rates of two neighboring limit orders crossed, an irrevocable trade was booked for both parties.

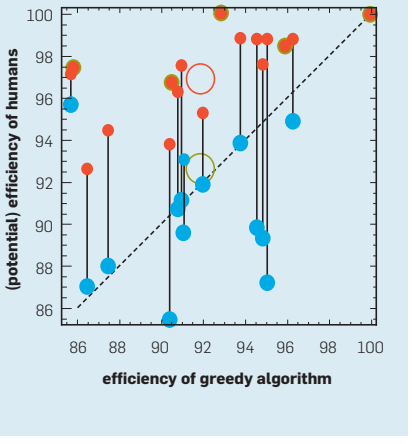
For the one-shot, simultaneous trade version of this model, there is a detailed equilibrium theory that precisely predicts the wealth of every player based on their position in the network;[11] in brief, the richest and poorest players at equilibrium are determined by finding the subset of vertices whose neighbor set yields the greatest contraction,[d] and this can be applied recursively to compute all equilibrium wealths. An implication is that the only bipartite networks in which there will not be variation in player wealths at equilibrium are those that contain perfect matchings. One of the primary goals of the experiment was to test this equilibrium theory behaviorally, particularly because equilibrium wealths are not determined by local structure alone, and thus might be challenging for human subjects to discover from only local interactions; even the best known centralized algorithm for computing equilibrium uses linear programming as a subroutine.[5] We again examined a wide variety of network structures, including several where equilibrium predictions have considerable variation in player wealth.

There were a number of notable findings regarding the comparison of subject behavior to the equilibrium theory. In particular, across all experiments and networks, there was strong *negative* correlation between the equilibrium predicted variation of wealth across players, and the collective earnings of the human subjects—even though there was strong *positive* correlation between equilibrium wealth variation and behavioral wealth variation. In other words, the greater the variation of wealth predicted by equilibrium, the greater the actual variation in behavioral wealth, but the more money that was left on the table by the subjects. This apparent distaste for unequal allocation of payoffs was confirmed by our best-fit model for player payoffs, which turned out to be a mixture of the equilibrium wealth distribution and the uniform distribution in approximately a (3/4; 1/4) weighting. Thus the equilibrium theory is definitely relevant, but is improved by tilting it toward greater equality. This can be viewed as a networked instance of inequality aversion, a bias that has been noted repeatedly in the behavioral game theory literature.[1]

Our experiments on networked bargaining[3] have a similarly financial flavor, and are also accompanied by an equilibrium theory.[4] In these experiments, each edge in the network represents a separate instance of Nash's bargaining game:[21] if by the end of the experiment, the two subjects on each end of an edge can agree on how to split $2, they each receive their negotiated share (otherwise they receive nothing for this edge). Subjects were thus simultaneously bargaining independently with multiple neighbors for multiple payoffs. Network effects can arise due to the fact that different players have different degrees and thus varying numbers of deals, thus affecting their "outside options" regarding any particular deal. In many experiments, the system also enforced limits on the number of deals a player could close; these limits were less than the player's degree, incentivizing subjects to shop around for the best deals in their neighborhood. The system provided a GUI that let players make and see separate counter offers with each of their neighbors.

Perhaps the most interesting finding regarding the comparison between subject performance and a simple

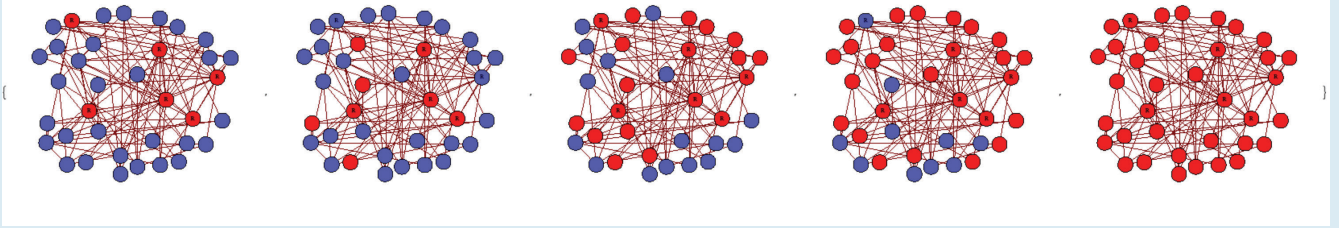
Figure 5. Human performance vs. greedy algorithm in networked bargaining, demonstrating the effects of subject obstinacy. Where occlusions occur, blue dots are slightly enlarged for visual clarity. The length of the vertical lines measure the significant effects of subject obstinacy on payoffs.



c As per the theoretical model, players were not able to offer different rates to different neighbors; thus conceptually prices label vertices, not edges.

d For instance, a set of 10 milk players who collectively have only three neighboring wheat players on the other side of the bipartite network has a contraction of 10/3.

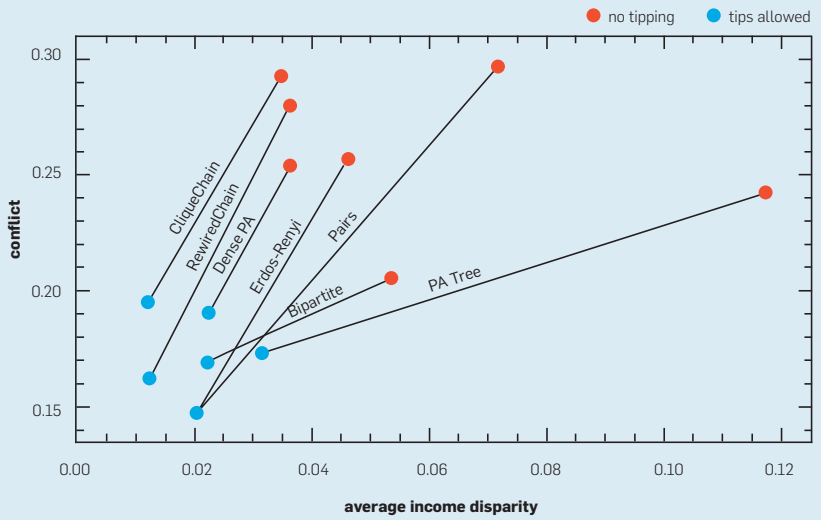
Figure 7. Series of snapshots of global state in a minority power biased voting experiment, showing an instance in which a minority player (upper left vertex R) acquiesces at various times though eventually wins out.



greedy algorithm for approximating the maximum social welfare solution, summarized in Figure 5. This centralized greedy algorithm simply selects random edges in the network on which to close bargains, subject to any deal limits in the experiment, until no further deals could be closed without violating some deal limit. The social welfare obtained (which does not require specifying how the edge deals are split between the two players) is then simply $2 times the number of closed deals, as it is for the behavioral experiments as well.

The blue dots in Figure 5 each represent averages over several trials of one of the network topologies examined (thus each dot corresponds to a different topological family). The *x* value shows the social welfare of the greedy algorithm as a percentage of the maximum social welfare (optimal) solution, while the y value shows the same measure for the human subjects. Averaged over all topologies, both humans and greedy perform rather well—roughly 92% of optimal (blue open circle). However, while the greedy solutions are maximal and thus cannot be locally improved, much of the inefficiency of the subjects can be attributed to what we might call the *Price of Obstinacy*: at the end of many experiments, there were a number of deals that still could have been closed given the deal limits on the two endpoints, but on which the two human subjects had not been able to agree to a split. If we simply apply the greedy algorithm to the final state of each behavioral experiment, and greedily close as many remaining deals as possible, the *potential* performance of the subjects on each topology, absent obstinacy, rises to the orange dot connected to the corresponding blue dot in the figure. This hypothetical subject performance is now well above the performance of pure greedy (all orange points above the diagonal now), and the average across topologies is close to 97% of optimal (orange open circle). In other words, the human subjects are consistently finding better underlying solutions than those obtained by simply running greedy on the initial graph, but are failing to realize those better solutions due to unclosed deals. While humans may show aversion to inequality of payoffs, they can also be stubborn to the point of significant lost payoffs.

Figure 6. From independent-set experiments: Average income disparity between neighbors (*x*-axis) vs. average time neighbors are conflicting kings (*y*-axis), both with (blue) and without (orange) side payments. Grouped by network structure. The side payments uniformly reduced conflict and disparity.



Independent set. Another set of experiments required subjects to declare their vertex to be either a "king" or a "pawn" at each moment, with the following resulting payoffs: any player who is the only one that has declared kingship in his neighborhood enjoys the highest possible rate of pay; but if one or more of their neighbors are also kings, the player receives nothing. On the other hand, pawns receive an intermediate rate of pay regardless of the states of their neighbors. It is easily seen that the Nash equilibria of the one-shot, simultaneous move version of this game are the maximal independent sets (corresponding to the kings) of the graph, while the maximum social welfare state is the largest independent set, whose centralized computation is *NP*-hard. Because we were concerned that computing payoffs based on only the final state of the gain

would lead to an uninteresting global "chicken" strategy (all players declaring king until the final seconds of the experiment, with some players then "blinking" and switching to pawn), in these experiments payoffs accrued continuously according to the prorated time players spent in each of the three possible states (pawn, king with no conflicting neighbors, conflicting kings).

Every experiment was run under two conditions—one just as described above, and another in which the GUI included an additional element: in the case that the player was the lone king in their neighborhood, and thus enjoying the highest rate of pay, a slider bar permitted them to specify a fraction of their earnings in that state to be shared equally among all their neighbors (whose pawn status allows the king's high payoff). These "tips" or side payments could range from 0% to 100% in increments of 10%, and could be adjusted at any time. Note that in some cases, depending on network structure, some vertices might be able to obtain a higher rate of pay by being a pawn receiving side payments from many neighboring kings than by being the lone king in their neighborhood.

The most striking finding was that, across a wide variety of network structures, the introduction of the side payments uniformly raised the collective payoffs or social welfare. Side payment rates were often generous, and averaged close to 20%. Furthermore, when side payments are introduced, both the average income disparity between neighboring players, and the amount of time they spend as conflicting kings, are considerably reduced, across all network structures examined (see Figure 6). This suggests that without side payments, subjects used conflict, which reduces the wealth of all players involved, to express perceived unfairness or inequality. The side payments reduce unfairness and consequently reduce conflict, thus facilitating coordination and raising the social welfare.

Biased voting. The biased voting experiments[14] shared with the earlier consensus experiments an incentive toward collective agreement and coordination, but with an important strategic twist. As in consensus, each player had to simply select a color for their vertex, but now only between the two colors red and blue. If within the allotted time, the *entire population* converged unanimously to either red or blue, the experiment was halted and every player received some payoff. If this did not occur within the allotted time, every player received nothing for that experiment. Thus the incentives were now not at the individual level, but at the collective—players had to not only agree with their neighbors, but with the entire network, even though they were still given only local views and interactions.

The side payments reduce unfairness and consequently reduce conflict, thus facilitating coordination and raising the social welfare.

The strategic twist was that different players were paid different amounts for convergence to the two colors within the same experiment. In particular, some players received a higher payoff for convergence to blue, while others received a higher payoff for convergence to red. Typical incentives might pay blue-preferring players $1.50 for blue convergence and only $0.50 for red, with red-preferring players receiving the reverse. Some experiments permitted asymmetries between higher and lower payoffs, thus incentivizing some players to "care" more about the color chosen by the population. These experiments thus set up a deliberate tension between competing individual preferences and the need for collective unity.

In the most dramatic set of experiments, networks were chosen according to preferential attachment—known to generate a small number of vertices with high degree—and the vast majority of players given incentives that paid more for convergence to blue. However, the minority of vertices preferring red was chosen to be the high-degree vertices. These experiments tested whether a small but well-connected minority could systematically impose its preferences on the majority, thus resulting in suboptimal social welfare.

The answer was resoundingly affirmative: in 27 such "minority power" experiments, 24 of them resulted in the subjects reaching a unanimous choice—in every case, the preferred choice of the well-connected minority. The finding is especially surprising when we remember that since everyone has only local views and information,

the powerful minority has no particular reason to believe they are powerful—in fact, their high degree ensured that at the start of each such experiment, they would see themselves surrounded by players choosing the opposing color. Indeed, the minority players would often acquiesce to the majority early in the experiment (see Figure 7, which shows a series of snapshots of actual play during an experiment). But the dynamics always eventually came to favor the minority choice.

A behavioral network formation game. Our most recent experiments[16] attempted to address what is perhaps the greatest of many artificialities in this line of research: the exogenous imposition of the social network structure mediating interactions. While corporations and other social entities of course often do impose organizational structure, it is natural to believe that in many circumstances, humans will organically construct the communication and interaction patterns required to solve a task efficiently—perhaps even circumventing any imposed hierarchy or structure. Given the aforementioned overall strong performance of our subjects across a wide variety of challenging tasks, even when network structures were complex and not directly optimized for the task, we were naturally interested in whether performance might improve even further if the subjects could collectively choose the networks themselves.

We thus ran among the first experiments in network formation games, on which there is an active theoretical literature.[8,24] We wanted to design such a game in which the formation of the network was not an end in itself, as it is in many of the theoretical works, but was in service of a collective task—which we again chose to be biased voting. The framework was thus as followed: the payoff functions for the players was exactly as described for biased voting, with all players wanting to reach unanimity, but having a preferred (higher payoff) color. Now, however, there were *no edges* in the network at the start of each experiment—every vertex was isolated, and players could thus see only their own color. Throughout the experiment, players could optionally and unilaterally *purchase* edges to other players, resulting in subsequent bilateral viewing of each other's colors for the two players; the GUI would adapt and grow each player's neighborhood view as edges were purchased. A player's edge purchases were deducted from any eventual payoffs from the biased voting task (subject to the constraints that net payoffs could never be negative).

These experiments thus tested whether a small but well-connected minority could systematically impose its preferences on the majority, resulting in suboptimal social welfare.

Players were thus doing two things at once—building the network by purchasing edges, and choosing colors in the biased voting task. The GUI had an edge purchasing panel that showed players icons indicating the degrees and shortest-path distances of players they were not currently connected to, thus allowing them to choose to buy edges (for instance) to players that were far away in the current network and with high degree, perhaps in the hopes that such players would aggregate information from distant areas of the network; or (for instance) to low-degree vertices, perhaps in the hope of strongly influencing them. The formation game adds to the biased voting problem the tension that while the players must collectively build enough edges to facilitate global communication and coordination, individual players would of course prefer that others purchase the edges.

While there were many detailed findings, the overall results were surprising: the collective performance on this task was by far the worse we have seen in all of the experiments to date, and much worse than on the original, exogenous network, biased voting experiments. Across all experiments (that included some in which the subjects started not with the empty network, but with some "seed" edges that were provided for free), the fraction in which unanimity was reached (and thus players received nonzero payoffs) was only 41%—far below the aforementioned nearly 90% efficiency across all previous experiments. We were sufficiently surprised that we ran control experiments in which a subsequent set of subjects were once again given fixed, exogenously imposed networks—but this time, the "hard" networks created by the network formation subjects in cases where they failed to solve the biased voting task. This was

done to investigate the possibility that the formation subjects built good networks for the task, but either ran out of time to reach unanimity, or included subjects who behaved very stubbornly because they had significant edge expenditures and thus strongly held out for their preferred color.

Performance on the control experiments was even worse. The surprising conclusion seems to be that despite the fact that subjects clearly understood the task, and were now given the opportunity to solve it not on an arbitrary network, but one collectively designed by the population in service of the task, they were unable to do so. One candidate for a structural property of the subject-built networks that might account for their difficulty in the biased voting task is *(betweenness) centrality*, a standard measure of a vertex's importance[e] in a network. Compared to the networks used in the original, exogenous-network biased voting experiments, the distribution (across vertices) of centrality in the subject-built networks is considerably more skewed.[16] This means that in the network formation experiments, there was effectively more reliance on a small number of high-centrality vertices or players, making performance less robust to stubbornness or other non-coordinating behaviors by these players. Indeed, there was moderately positive and highly significant correlation between centrality and earnings, indicating that players with high centrality tended to use their position for financial gain rather than global coordination and information aggregation.

Despite their demonstrated ability to solve a diverse range of computational problems on a diverse set of networks, human subjects seem poor at *building* networks, at least within the limited confines of our experiments so far. Further investigation of this phenomenon is clearly warranted.

Concluding Remarks

Despite their diversity, our experiments have established a number of rather consistent facts. At least in moderate population sizes, human subjects can perform a computationally wide range of tasks from only local interaction. Network structure has strong but task-dependent effects. Notions of social fairness and inequality play important roles, despite the anonymity of our networked setting. Behavioral traits of individual subjects are revealed despite the highly simplified and stylized interactions; with language removed, subjects persistently try to invent signaling mechanisms.

There are a number of recent efforts related to the research described here. Some compelling new coloring experiments[7,20] have investigated the conditions under which increased connectivity improves performance. Our experimental approach has thus far aimed for breadth, but studies such as these are necessary to gain depth of understanding. We have also usually done only the most basic statistical analyses of our data, but others have begun to attempt more sophisticated models.[6]

Perhaps the greatest next frontier is to conduct similar experiments on the Web, where a necessary loss of control over subjects and the experimental environment may be compensated by orders of magnitude greater scale, both in population size and the number of experimental conditions investigated. Recent efforts using both the open web and Amazon's Mechanical Turk online labor market have started down this important path.[2,19,23]

Acknowledgments

Many thanks to the stellar colleagues who have been my coauthors on the various papers summarized here: Tanmoy Chakraborty, Stephen Judd, Nick Montfort, Sid Suri, Jinsong Tan, Jennifer Wortman Vaughan, and Eugene Vorobeychik. I give especially warm acknowledgments to Stephen Judd, who has been my primary collaborator throughout the project. Thanks also to Colin Camerer and Duncan Watts, who both encouraged me to start and continue this line of work, and who made a number of important conceptual and methodological suggestions along the way.

e The betweenness centrality of vertex v is average, over all pairs of other vertices u and w, of the fraction of shortest paths between u and w in which v appears.

References

1. Camerer, C. *Behavioral Game Theory*. Princeton University Press, Princeton, NJ, 2003.
2. Centola, D. The spread of behavior in an online social network experiment. *Science 329*, 5996 (2010), 1194–1197.
3. Chakraborty, T., Judd, J.S., Kearns, M., and Tan, J. A behavioral study of bargaining in social networks. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2010, 243–252.
4. Chakraborty, T., Kearns, M., and Khanna, S. Networked bargaining: Algorithms and structural results. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2009,159–168.
5. Devanur, N.R., Papadimitriou, C.H., Saberi, A., and Vazirani, V.V. Market equilibrium via a primal-dual algorithm for a convex program. *Journal of the ACM 55*, 5 (2008).
6. Duong, Q., Wellman, M.P., Singh, S., and Kearns, M. Learning and predicting dynamic behavior with graphical multiagent models. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems* (2012).
7. Enemark, D.P., McCubbins, M.D., Paturi, R., and Weller, N. Does more connectivity help groups to solve social problems? In *ACM Conference on Electronic Commerce*. ACM Press, New York, 2011, 21–26.
8. Jackson, M.O. A survey of models of network formation: stability and efficiency. In *Group Formation in Economics: Networks, Clubs and Coalitions*, Cambridge University Press, 2005.
9. Jackson, M.O. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, 2010.
10. Judd, S., Kearns, M., and Vorobeychik, Y. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences 107*, 34 (2010), 14978–14982.
11. Kakade, S.M., Kearns, M.J., Ortiz, L.E., Pemantle, R., and Suri, S. Economic properties of social networks. *Neural Information Processing Systems* (2004).
12. Kearns, M. Networked Life. University of Pennsylvania Undergraduate Course; http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife
13. Kearns, M. and Judd, J.S. Behavioral experiments in networked trade. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2008, 150–159.
14. Kearns, M., Judd, S., Tan, J., and Wortman, J. Behavioral experiments in biased voting in networks. *Proceedings of the National Academy of Sciences 106*, 5 (2009), 1347–1352.
15. Kearns, M., Judd, S., and Vorobeychik, Y. Behavioral conflict and fairness in social networks. *Workshop on Internet and Network Economics* (2011).
16. Kearns, M., Judd, S., and Vorobeychik, Y. Behavioral experiments on a network formation game. *ACM Conference on Electronic Commerce*. ACM Press, New York, 2012.
17. Kearns, M., Suri, S., and Montfort, N. An experimental study of the coloring problem on human subject networks. *Science 313* (2006), 824–827.
18. Khatib, F., Cooper, S., Tyka, M.D., Xu, K., Makedon, I., Popovic, Z., Baker, D., and Foldit Players. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* (2011).
19. Mason, W. and Suri, S. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* (2011).
20. McCubbins, M.D., Paturi, R., and Weller, N. Connected coordination network structure and group coordination. *American Politics Research 37*, 5 (2009), 899–920.
21. Nash, J.F. The bargaining problem. *Econometrica 18*, 2 (1950), 155–162.
22. Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V.V., Eds. *Algorithmic Game Theory*. Cambridge University Press, 2007.
23. Suri, S. and Watts, D.J. Cooperation and contagion in Web-based, networked public goods experiments. *PLoS ONE 6*, 3 (2011).
24. Tardos, E. and Wexler, T. Network formation games and the potential function method. In *Algorithmic Game Theory*, Cambridge University Press, 2007, 487–513.
25. Watts, D.J. *Six Degrees: The Science of a Connected Age*. W.W. Norton & Company, 2003.
26. Wikipedia. Crowdsourcing; http://en.wikipedia.org/wiki/Crowdsourcing en.wikipedia.org/wiki/Crowdsourcing.

Michael Kearns (mkearns@cis.upenn.edu) is a professor in the Computer and Information Science Department of the University of Pennsylvania. His research interests include machine learning, social networks, algorithmic game theory, and computational finance.

Behavioral Experiments on a Network Formation Game

Michael Kearns, University of Pennsylvania
Stephen Judd, University of Pennsylvania
Yevgeniy Vorobeychik, Sandia National Laboratories

Abstract: We report on an extensive series of behavioral experiments in which 36 human subjects collectively build a communication network over which they must solve a competitive coordination task for monetary compensation. There is a cost for creating network links, thus creating a tension between link expenditures and collective and individual incentives. Our most striking finding is the poor performance of the subjects, especially compared to our long series of prior experiments. We demonstrate that the subjects built difficult networks for the coordination task, and compare the structural properties of the built networks to standard generative models of social networks. We also provide extensive analysis of the individual and collective behavior of the subjects, including free riding and factors influencing edge purchasing decisions.

Categories and Subject Descriptors: Computer Applications [**Social and Behavioral Sciences**]: Economics, Psychology, Sociology

General Terms: Economics, Experimentation, Human Factors, Theory

Additional Key Words and Phrases: Social Networks, Game Theory, Network Formation

1. INTRODUCTION

In recent years, research from a variety of disciplines has established the universality of certain approximate structural properties of large-scale social, technological, organizational and economic networks. These properties include networks having small diameter, high clustering of connectivity, and heavy-tailed degree distributions. The apparent ubiquity of these properties, despite the diversity of the domains of the networks in which they appear, has led researchers to seek explanations in the form of models of network formation that can reliably generate the observed structures.

The most studied class of such models are *stochastic* network formation models, in which networks form through a decentralized process that generates local connectivity using randomization; examples include the classic Erdös-Renyi random graph model [Bollabas 2001], and the more recent small worlds [Watts and Strogatz 1998; Kleinberg 2000] and preferential attachment [Barabasi and Albert 1999] models. These models have been successful in providing simple and relatively general mechanisms generating common structural properties of large networks.

An important criticism of the stochastic formation models is that in real networks, connectivity rarely forms entirely randomly — rather, there is often some significant component of *purposefulness* when a new node or link is formed. Professionals join and form connections on a service like LinkedIn to enjoy the career benefits of being part of that network; new web sites are created to generate traffic, including by being

---

This work was conducted while Y.Vorobeychik was at the University of Pennsylvania.

EC'12, June 4–8, 2012, Valencia, Spain.

part of the network of links and the attendant search indexing benefits. While there may be elements of arbitrariness or stochasticity, networks generally arise from the self-interests of their constituents, and serve some collective purpose(s).

An alternative class of economic or game-theoretic models directly addresses the issue of self-interest and purpose in network growth. In *network formation games* [Tardos and Wexler 2007; Jackson 2005; 2010; Fabrikant et al. 2003; Borgs et al. 2011; Albers et al. 2006; Brautbar and Kearns 2011; Even-Dar et al. 2007; Even-Dar and Kearns 2006], individual players typically have utility functions with two competing components: there is a cost to join the network, usually in the form of purchasing links to other players (where the cost may be viewed as monetary, as in the connectivity and physical costs of adding a router to the Internet, or more cognitive, as in the time needed to create and maintain friendships on Facebook). But after joining the network, a player enjoys participation benefits (perhaps abstracted by some measure of their centrality in the network), and their overall payoff is the network benefit minus their cost of joining. It is common to equate the outcome of such games with their Nash equilibria, just as the stochastic models are analyzed for their statistically typical properties. While there has been growing interest in the theory of network formation games for several years now, to our knowledge there is not an accompanying behavioral literature.

In this paper, we describe among the first and largest human-subject experiments in a pure network formation game. These are the most recent in a long series of behavioral experiments on strategic and financial interactions in social networks [Kearns et al. 2006; Kearns et al. 2009; Judd et al. 2010; Kearns and Judd 2008; Chakraborty et al. 2010; Kearns et al. 2011], but represent a major departure from our prior experiments, where the networks examined were always exogenously imposed on the subjects. Here we *endogenized* the formation of the network structure itself as part of the experiment. While the theoretical literature on network formation games has focused on one-shot, simultaneous move games of full information (and even there, characterizations of equilibria are difficult and elusive), our experiments investigate a formation game of continuous, asynchronous moves and partial information.

In the experiments, subjects were given financial incentives to solve a collective but competitive coordination problem of *biased voting*, in which they must unanimously agree on one of two alternative choices, or receive no payoff at all. The competitive aspect arises from the fact that different players have different financial preferences for which of the two choices is agreed upon. We have previously studied this problem on fixed network structures [Kearns et al. 2009]; in the current experiments the subjects themselves had to *build* the network during the experiment, via individual players purchasing links whose cost is subtracted from their eventual task payoff. The nature of the biased voting task and the financial self-interests of the players sets up a clean strategic tension: in order to solve the biased voting problem, the players must collectively purchase enough links to establish some minimal global connectivity; but any individual player would prefer others to incur the costs of building this shared infrastructure.

A striking finding is that the players performed very poorly compared to our long series of prior experiments in which network structures were imposed exogenously. Despite clearly understanding the biased voting task, and being permitted to collectively build a network structure facilitating its solution, subjects instead appear to have built very *difficult* networks for the task. This finding is in contrast to intuition, case studies and theories suggesting that humans will often organically build communication networks optimized for the tasks they are charged with, even if it means overriding more hierarchical and institutional structures [Burns and Stalker 1994; Nishiguchi and Beaudet 2000]. We also report on a number of other aspects of subject behavior

and performance, including structural properties of the built networks, comparisons to standard network formation models, and free-riding in edge purchasing.

2. EXPERIMENTAL DESIGN AND METHODOLOGY

The experiments reported here were held in three sessions with different pools of 36 subjects each. As in our previous experiments, subjects sat at networked workstations separated by physical partitions, and the only communication permitted was through the system. In each of many short experiments, subjects were given financial incentives to solve a global coordination problem via only local interactions in a network. In prior experiments, the network structure was a design variable that we chose and imposed exogenously in each experiment; here, the network structure was created by the subjects themselves, as described below.

We first describe the overarching collective task the subjects were charged with solving. In each experiment, subjects sat at individual workstations, and each controlled the state of a single vertex in a 36-vertex network whose connectivity structure evolved throughout the experiment. The state of a subject's vertex was simply one of two colors (red or blue), and could be asynchronously updated as often as desired during the one-minute experiment. Subjects were able to view the current color choices of *only their immediate neighbors in the current network* at all times. No communication between subjects outside the experimental platform was permitted.

In each experiment, each subject was given a financial incentive that varied across the population, and specified both individual preferences and the demand for collective unity. For instance, one player might be paid $2 for blue consensus and $1 for red consensus, while another might be paid $1 for blue consensus and $2 for red consensus, thus creating distinct and competing preferences across individuals. However, payments for an experiment were made only if (red or blue) *global* unanimity of color was reached; thus subjects had to balance their preference for their higher payoff color with their desire for any payoff at all.

At the beginning of each 1-minute experiment, the network over the players was typically *empty*: there were no edges, and thus every player controlled an isolated vertex and could see only their own color choice. Clearly reaching unanimity of color choice in the biased voting task is highly unlikely in such circumstances. Thus at any time throughout an experiment, subjects were free to *purchase* edges to other players at a fixed cost. Edge purchases were unilateral — they did not require approval by the player on the receiving end — but their benefits were bilateral, meaning that after the purchase both players could see each other's current color choices. The system GUI (see Figure 1) would dynamically evolve as new edges were purchasing, always showing a subject the color choices of their neighbors in the current network.

An important design decision is what information the players are provided to help them decide *which* vertices to purchase edges to. One possibility is *no* information: players could simply indicate their desire for a new connection, and the system could simply give them a new edge to a random player to whom they were not already connected. However, this would predetermine the network topologies built to be of a random, unstructured, Erdös-Renyi variety, and not particularly useful for the biased voting task. We thus decided to give subjects two pieces of information about their current non-neighbors: their current *degree* (number of connections), and their current *shortest-path distance* in the network from the subject. This allowed players to selectively purchase edges that seem relevant to the task. For instance, players could choose to buy edges to players with high degree who were distant from them in the current network (perhaps in the hopes that such players aggregate information on the other side of the network), or to players with zero or low degree (perhaps in the hopes of having strong influence over the color choice of such players). We emphasize that

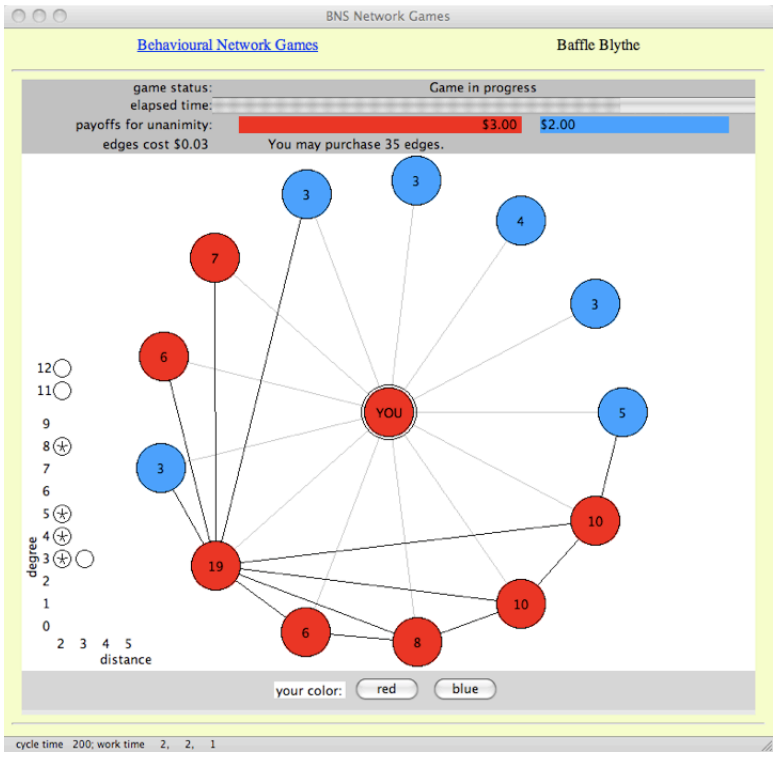


Fig. 1. Sample screenshot of player GUI in the network formation game. Around the vertex labeled "YOU", the central panel displays the player's current network neighborhood, indicating the current color of their neighbors as well as any edges between neighbors. In the action panel at the bottom, the player can change their own current color by clicking on the buttons labeled "red" and "blue". To the lower left of the central panel is the grid where players can select other vertices to purchase edges to. Each non-neighbor is represented by a circle whose grid position indicates their current degree and shortest-path distance from the player. (Vertices not currently in the same connected component as the player are shown as being at infinite distance.) If more than one vertex has the same degree and distance, the circle contains a "*" symbol. The player purchases edges by clicking on the desired circle, at which point their new neighbor will be incorporated into the neighborhood display. At the top, the time elapsed in the experiment is shown, along with the payoffs of the player, which are dynamically reduced as edges are purchased. The fixed cost per edge is also displayed.

this choice of informational design was not made with realism and generality in mind — obviously, in real social networks one does not have knowledge of the degree and distance of non-neighboring vertices — but rather potential relevance to the collective task, which seemed to us a more important experimental criterion.

We also note that this informational design also permitted the subjects, in principle, to collectively generate networks similar to those of well-studied stochastic models such as Erdös-Renyi (by simply ignoring the degree and distance values, and always choosing a random non-neighbor to connect to), or networks similar to those generated by preferential attachment (by ignoring distance information, and favoring purchasing edges to higher-degree vertices). We know from previous fixed-network experiments that such models generate networks generally favorable for human performance across

a wide variety of tasks, including biased voting [Kearns et al. 2009]. Thus at least some behaviorally "easy" networks are collectively reachable within the given system design.

Each player's GUI had an edge-purchase panel in which each of their current non-neighbors was represented by a circle on a 2-dimensional grid, indicating that non-neighbor's current degree and distance from the player; see Figure 1. By simply clicking on the corresponding circle, a player would purchase an edge, and the new neighbor and their color would be dynamically incorporated into their network neighborhood display, and remain for the duration of the experiment. (Edge purchases were persistent and irrevocable.) If more than one non-neighbor had the same degree and distance, the grid circle would indicate so. As an experiment progressed, degrees increased and distances decreased in the growing network.

If the players failed to reach unanimity in the allotted time, all edge purchases were forgiven, and no payoffs were made; but if unanimity was reached at any point, the experiment was terminated, and a player's edge purchases were subtracted from their earnings on the biased voting problem to arrive at their net payoff. The system also enforced the condition that players must have strictly positive payoffs on successful experiments: thus, each player could only spend an amount on edges that was slightly and strictly *less* than their lower-payoff color in the biased voting problem. This prevents players from becoming "infinitely stubborn" in favor of their higher-payoff color if their lower payoff has been reduced to zero by edge purchases.

In a subset of experiments, conditions were as described above, but instead of starting with the empty network (which we shall refer to as *unseeded* experiments in the sequel), the experiment began with a "seed" network of edges that were provided free of charge to the players [Kleinberg 2000; Even-Dar and Kearns 2006]; players could then optionally purchase additional edges as above. Thus each experiment was characterized by the distribution of biased voting incentives of the players, the presence or absence of the seed network and its structure, and the fixed price of edges (which we varied from experiment to experiment). We shall comment on each of these design variables in the appropriate places as we describe our findings.

From a theoretical perspective, we thus presented our subjects with a task-oriented network formation game of partial information (unknown incentives or types of the other players, unknown and evolving global network structure) and asychronous, repeated moves with finite termination time. We note that formal analysis of even vastly simplified versions of this game appears to be quite challenging, but might be an interesting avenue for future work.

3. BACKGROUND ON PRIOR FIXED-NETWORK EXPERIMENTS

As mentioned above, we have conducted experiments similar to those described here since 2005, but always designing and exogenously imposing the network structures mediating interaction. The tasks we have given to subjects are diverse, and include graph coloring [Kearns et al. 2006], consensus [Judd et al. 2010], networked trading [Kearns and Judd 2008], networked bargaining [Chakraborty et al. 2010], independent set [Kearns et al. 2011], and biased voting [Kearns et al. 2009]. While direct comparisons across tasks can be difficult, there is one general and easily measured metric of collective performance, which is the *efficiency:* in any given experiment, we can compute the configuration of play that would have maximized the total payments to the subjects, and then compute the fraction of that maximum payoff the population actually realized. We can then average this quantity across all experiments ever conducted, regardless of task, network structure, and other design variables. The resulting value is 0.88 — in other words, over the lifetime of the project, subjects have extracted almost 90% of the value that was available to them in principle. We conclude

that humans are quite good at solving a variety of challenging tasks from only local interactions in an underlying network.

In the previous biased voting experiments on exogenous networks, 55 of 81 experiments resulted in unanimity and therefore some payoff to all subjects, again giving fairly strong collective performance. On the subset of experiments in which the network and incentives had what we called a *minority power* structure, performance was even stronger, with 24 of 27 experiments reaching consensus. We shall contrast these findings with those for biased voting with network formation.

4. RESULTS

4.1. Overall Performance

Our experiments were structured in three separate sessions with different subject pools; while the conditions in the first two sessions were similar and designed in advance, the third session was designed in response to the findings of the first two, and shall be discussed separately below.

Session 1 consisted of 99 short experiments, with 63 of these being unseeded; Session 2 consisted of 72 experiments, 27 of which were unseeded. Across these 171 experiments, various other conditions varied as well, including the cost per edge, the fraction of players with higher payoffs for red, and the relative strengths of the incentives for players of different types. We shall discuss the effects of these design variables later, and for now focus on the collective performance across all these conditions in Sessions 1 and 2.

Compared to our long series of prior fixed-network experiments, that performance was surprisingly poor: Session 1 produced only 47% successful (unanimous) outcomes, and Session 2 only 39%, for an overall success rate of 44%. This is in sharp contrast to the aforementioned efficiency across all tasks of 88% — approximately double that of the current experiments — and the 68% success rate of the fixed-network biased voting experiments, more than 20% higher than in the network formation game. It appears that allowing the subjects to control the creation of the network significantly *worsened* collective performance [1].

There are at least two plausible explanations for this degradation in performance that do not simply entail that subjects built "bad" networks for the task. The first explanation is one of cognitive *overload*: perhaps subjects built "good" networks, but simply ran out of time to solve the biased voting problem on those good networks. The second explanation is one of *stubbornness* due to modified incentives: perhaps the subjects built good networks, but due to the edge purchases, some players had reduced the net payoff of their less preferred color to such a small amount that they might be very resistant to acquiesce to the majority color, resulting in stalemates. This hypothesis is made more plausible by the significant amount of "free riding" that occurred with respect to edge purchases, discussed later.

To investigate the overload and stubbornness hypotheses, we designed and conducted a third session of experiments with fresh subjects. In Session 3, each game was seeded with a network that was the *final* network constructed by the subjects in an unseeded Session 1 experiment. In these Session 3 experiment, the subjects *only* played the biased voting game — *no edge purchases* were allowed by the system. The subjects were thus back in the setting of our earlier, exogenous fixed-network experiments, but this time using networks *built by previous human subjects*. We deliberately chose a subset of the final networks from Session 1 on which the performance there was particularly poor — namely, 18 final networks on which the success rate was only

[1] In the games that failed to converge, the average size of the minority was actually smaller than it was for the failed biased voting games, but not to the level of statistical significance.

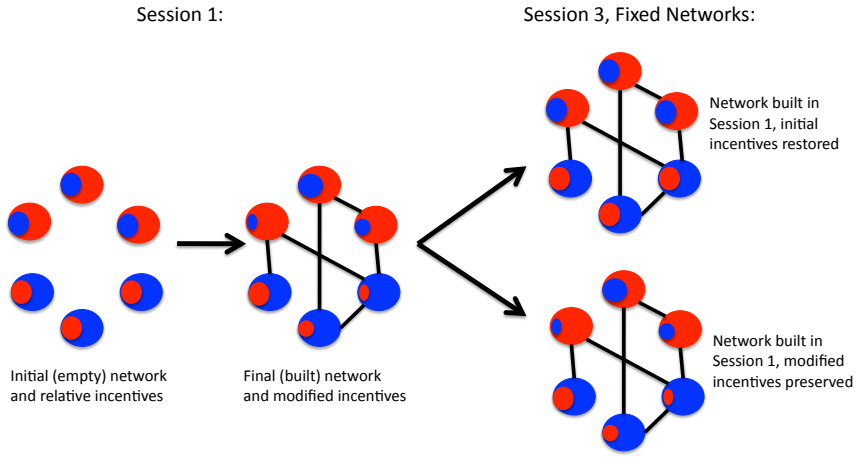


Fig. 2. Design of Session 3 experiments. At the start of an unseeded Session 1 experiment (left), the network is empty and some players (the upper 3) have higher payoffs for red than blue, schematically represented by the ratio of the red to blue areas in each vertex. At the completion of the experiment (middle), a network has been built by the players, and some players may have more extreme relative preferences due to the reduction of their payoffs by edge purchases. In Session 3 (right), we took the final networks built by Session 1 experiments, and exogenously imposed them as the networks of a pure biased voting game without any edge purchases. We did so using both the original, restored incentives of Session 1, and the post-edges modified incentives.

17%. We ran each Session 3 network under two different incentive conditions: one in which the incentives the players were the same as at the *beginning* of the corresponding Session 1 experiment, and one in which they were the same as at the *end* (after edge expenditures). See Figure 2 for a description of Session 3 design.

Together these conditions allow us to investigate the validity of the explanations above: if subjects were simply running out of time in Session 1 (the overload hypothesis), they should fare much better in Session 3, since now the network formation task is removed, and they can focus only on the biased voting task; and if the difficulty in Session 1 was due to stubbornness after edge purchases, the Session 3 experiments in which the Session 1 networks are used but the incentives are restored to their starting values should be more successful.

The stubbornness and overload hypotheses are strongly refuted by the Session 3 results. Success rates were slightly higher than in Session 1, but not significantly so. The strong signal was that Session 3 success rates in games using the original payoffs, and in games using residual payoffs, were both significantly lower than in our earlier fixed-network biased voting games [Kearns et al. 2009], with $P < 0.01$ in both cases. The success rate (57/81) during all games in the fixed-network biased voting session was significantly higher than that of all games (100/243) in the three network formation sessions ($P < 0.0001$). More pointedly, the games in each of the three network formation sessions are individually lower than the fixed-network biased voting games (all with $P < 0.01$).

We are thus led to the conclusion that our subjects *built networks on which it was simply difficult to accomplish the very task they were being paid to accomplish*. They appear to have had enough time and incentive, but built inherently poor networks. As per the earlier discussion, this is the first task of the many we have investigated in which human performance was so low.

4.2. Effects of Seed Networks

Recall that a subset of the Session 1 and 2 experiments explored network formation in which the subjects were provided an initial seed network as free infrastructure. Our goal was to examine whether having this seed network, which might facilitate communication and coordination at no cost, would allow the players to build better networks and yield stronger performance.

We examined three types of seed network structure: a 2-dimensional grid or torus network, which provides global connectivity with relatively few edges; a network of 6 cliques of size 6, which groups the players into small highly-connected communities; and preferential attachment networks that were also the focus of the minority power experiments we discuss shortly. Again, in each of these experiments the players were free to purchase additional edges that would be dynamically added to the seed network.

Again somewhat surprisingly, none of these seed network structures seemed to improve collective performance much, with the completion rates on seed torus experiments being 33% (6/18), and on seed clique experiments being 33% (9/27) — neither of which is significantly different from the unseeded networks of Session 1, but both of which are significantly lower for the previous fixed-network biased voting networks ($P < 0.01$). Whatever network the subjects were given to start, they seem to have turned it into a poor network for the task.

A major finding of our original biased voting experiments focused on experiments in which networks were generated according to preferential attachment, which results in a heavy-tailed degree distribution, and where we gave a (sometimes very small) minority of the players a higher payoff for red, with the majority preferring blue. The twist was that the red minority consisted of the highest-degree vertices in the network; we were thus investigating whether a small but well-connected minority could systematically impose its preference against the majority's. The answer was resoundingly positive: 24 of 27 (89%) such fixed-network experiments ended in consensus, every one of them on the minority preference [Kearns et al. 2009].

In the current experiments, we were interested in how this finding might be changed if subjects could add edges to the minority power networks. We thus ran a number of experiments in which both the seed network structure and arrangement of incentives were identical to those in the earlier minority power experiments. The success rate was 61% (22/36) — higher than for torus or clique seeds, but still much lower than the original exogenous network minority power rate of 89%. Once again, permitting the subjects to modify the network has harmed collective performance. However, now 35% of the successful experiments ended with the *majority* preference — compared with none in the exogenous network case, a dramatic change. One interpretation is that permitting the purchase of edges allows the majority players to better realize they are in the majority — which may have been difficult in the exogenous network case, especially for the preponderance of low-degree vertices — and causes them acquiesce to the minority less readily. This could account both for the lower overall success rate, and the increased rate of majority victories.

4.3. Effects of Edge Costs and Incentives

Recall that across the many experiments, we varied the cost per edge purchase, and the absolute and relative incentives across the population. Edge costs were low ($0.01), medium ($0.10), or high ($0.25), with the proviso that the edge purchases of a player must always be strictly less than $1 (which was always the payoff of the less preferred color).

Not surprisingly, the cost per edge had a strong effect on the resulting network density (as we shall see in the following section), but also on collective performance. The overall success rates for unseeded Session 1 experiments were 67% for low-cost experiments, 38% for medium-cost experiments, and 14% for high-cost experiments; the differences between these quantities are all significant at $P < 0.05$. We note that although there was a clear relationship between edge costs and performance, with higher costs resulting in worse performance, in no case did the subjects collectively approach the maximum allowed edge expenditures; the fraction of possible edge purchases in unseeded Session 1 experiments was 64% for low edge cost, 42% for medium, and 59% for high. Thus subjects could have built considerably denser networks in all cases, but chose not to. We also note that the seeded experiments provided the subjects with considerably denser networks for less edge expenditures, yet failed to significantly improve performance.

The incentives given to players also had a pronounced effect on subject performance. Recall that in each experiment a subject always desired unanimity (no payment was distributed unless unanimity was reached), but had a preference for one color over another. For example, a subject might receive 4 if all chose blue, and only 1 if the consensus was to red. We maintained the smaller incentive at 1 for all experiments, and varied the payoff of the preferred color, from 4 (strong preferences), 2 (weak preferences), and 1 (indifference between the color choices). The overall success rates for Session 1 experiments were 58% (7/12) when the subjects were indifferent between the two colors, 53% (19/30) when they had a weak preference for one of the colors, and only 17% (4/24) when they had a strong color preference. While there was no statistically detectable difference between the impact of weak and no preference, strong preference had a clear detrimental effect on solution rate ($P < 0.01$, comparing against weak preference).

4.4. Network Structure and Centrality Skewness

Both as a general matter, but especially in light of the overall poor performance of the subjects, the structure of the networks built during the experiments is a topic of central interest. How do the built networks compare to more naturally evolved social networks, and what properties of the built networks might account for the difficulty they posed for the biased voting problem? Here we initially restrict our analysis to the majority of experiments where there was no seed network, so that all structure was built by the subjects themselves.

We begin by establishing that the built networks actually do share a number of structural "universals" that appear frequently in real-world networks. The first remark worth making is that in every unseeded experiment, the subjects built a connected network — there were never two or more disconnected components. Thus regardless of the edge costs and other parameters, the subjects always bought enough edges to establish global communication.

The diameters (pairwise average shortest path distances) of the built networks were generally quite small compared to the population size; while the diameters show a strong dependence on the network density and therefore the edge costs, they averaged 1.32 for all low-cost experiments (standard deviation 0.17), 1.87 (standard deviation

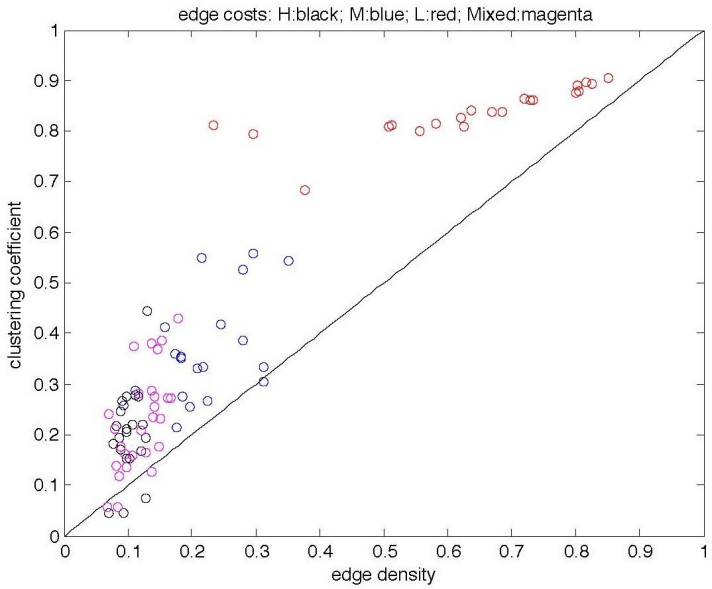


Fig. 3. Clustering coefficients vs. edge density for networks built in all unseeded experiments. For each built network, the x value indicates the fraction of possible edges present, and the y value indicates the clustering coefficient of the network. Random networks of the same densities would have clustering coefficients equal to their density, as suggested by the diagonal line. We see that clustering in the built networks is uniformly higher except at very low densities, where presumably the subjects may be been primarily concerned with establishing global connectivity. There are strong effects of edge costs (coded by color); higher edge costs consistently lead to lower densities and clustering. The mixed experiments had variable edge costs for the players, but always either medium or high.

0.19) in medium-cost experiments, and 2.38 (standard deviation 0.16) in high-cost experiments. Also as is for typical social networks, the clustering coefficients of the built networks were generally much higher than for random networks of the same density; see Figure 3. Furthermore, examination of the degree distributions of the built networks reveals the presence of "connector" vertices whose degrees are several times larger than the mean, another commonly cited property of natural social networks.

Given that the structural properties cited so far were generally present in our earlier, fixed-network experiments — in which the subjects performed much better — what can account for the difficulty of the built networks? While it is impossible to be certain, due to the complexities of both the networks and subject behavior, a strong candidate is the overreliance on very few vertices or subjects for connectivity and communication. In Figure 4 we demonstrate this reliance visually by showing, for three of the built networks, the effects on structure and connectivity of deleting a few vertices with the highest degrees. In each case, the network quickly becomes highly fragmented, a property generally true of the built networks.

We can make this analysis more systematic and rigorous by considering the quantity known as *betweenness centrality* (centrality in the sequel). Centrality is designed to measure, for each vertex u, the extent to which the rest of the network relies on u for its global connectivity and communication. More formally, we define

$$C_B(u) = \sum_{v,w \in V : v \neq u, w \neq u, v \neq w} \frac{n^u_{v,w}}{n_{v,w}}$$

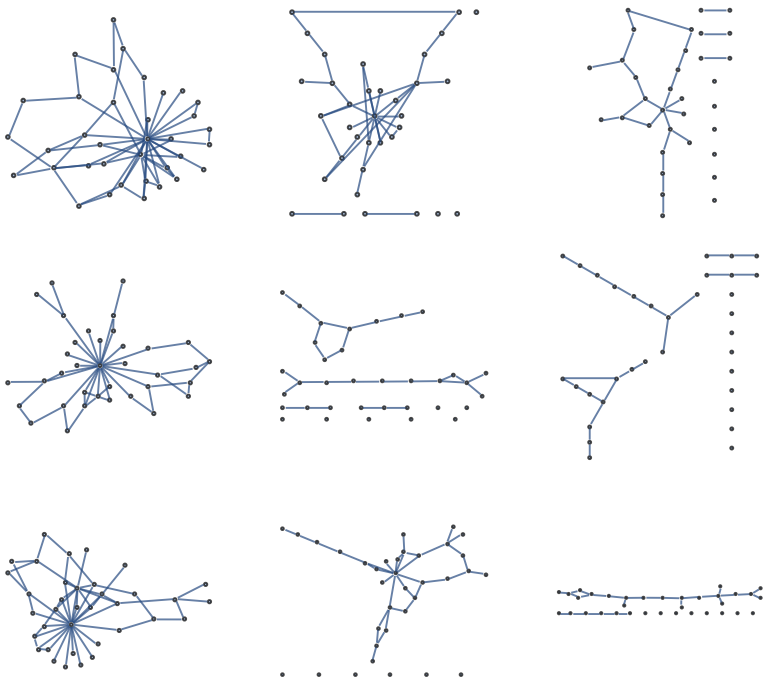


Fig. 4. Visualization of built networks in three unseeded Session 1 experiments with high edge costs. The first column shows a visualization of each of the built networks. Subsequent columns show the networks after repeated deletions of the highest-degree vertex remaining. In each case, the first deletion already shatters the network into multiple connected components, and subsequent deletions yield a large number of isolated vertices.

where V is the set of all vertices, $n_{v,w}$ is the number of shortest paths between v and w, and $n^u_{v,w}$ is the number of shortest paths between v and w that pass through vertex u. Thus $C_B(u)$ is a global measure of how often vertex u appears on shortest paths between all pairs of other vertices; it is a common metric of influence on communication and connectivity in social networks.

Echoing the analyses above, it turns out that the subject-built networks systematically differ from naturally occurring networks, and the ones we imposed on subjects in our earlier experiments, in their distribution of centrality. In particular, as with degrees, the built networks display a considerably more *skewed distribution* of $C_B(u)$: compared to natural network models at the same edge density, there are more vertices with very high and very low C_B , and fewer with intermediate values of C_B. See Figure 5.

There are a number of obvious reasons why overreliance on a few high-centrality vertices might make the biased voting task difficult. If a large fraction of the population implicitly relies on high centrality vertices to be effective aggregators of global information (such as the current majority color), noisy or selfish behavior by these individuals can impede collective performance. In the successful Session 1 experiments, the correlation between centrality and whether a subject received their higher payoff was both positive (0.18) and highly significant ($P < 0.001$), suggesting that high-centrality players may have implicitly used their position to influence outcome rather

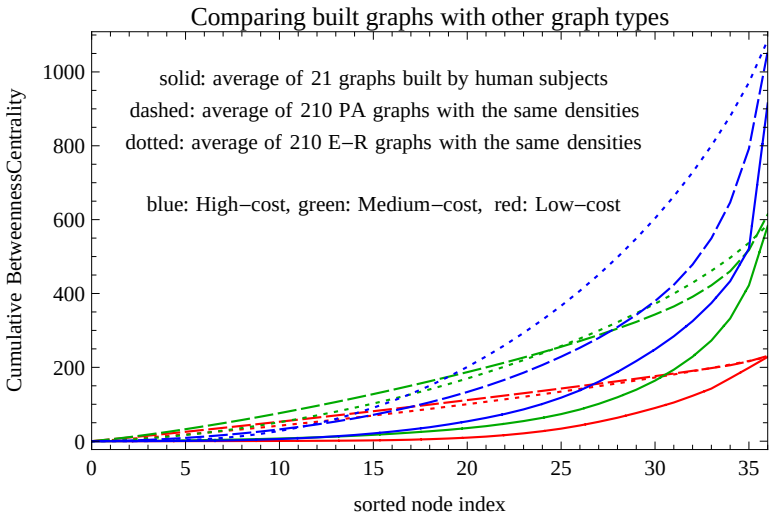


Fig. 5. Comparison of the distribution of centrality C_B between subject-built networks and standard generative models. For a given network, we sort vertices in order of increasing C_B values, and then compute the sum of all C_B values through a given rank in the ordering. We then average such curves over many built networks or many sampled networks from the generative models. We group the averaged curves by edge costs (color coded in the figure) for the built networks, and compare them to Erdös-Renyi and preferential attachment networks of the same average density. At each edge cost, we see that the built networks (solid lines) have much more skewed distributions of C_B than Erdös-Renyi (dotted lines) and preferential attachment (dashed lines): while the sum of all C_B values (rightmost point) is comparable for all three classes of networks at each density, much more of the cumulative centrality is accounted for by the final few, most central, vertices in the built networks, whose curves are considerably more convex than for the models.

than coordinate behavior — potentially contributing to the great majority of failed games.

We note that one might be tempted to think that the starting seeds would dissipate the propensity for skewness in the C_B values. For instance, in the torus networks, there already are 4 edges uniformly assigned to each node, so one could imagine new edges would have less of a biasing effect than when starting from an empty network. However, this intuition is misleading. When we examined the C_B values in networks that were seeded, we found final values that were both higher and lower than the starting values. The variance in people's buying behavior injected a variance into the C_B values, and they were spread out in both directions.

4.5. Purchasing Behavior and Free Riding

Thus far we have offered evidence that subjects built poor networks for the given task. It is natural to ask what particulars of subject behavior accounted for this. In this section, we examine the distribution of edge purchases in the population, which sheds some light on this question. Here the most striking aspect is the preponderance of free-riding: at all edge costs and in each experiment, there is a significant fraction (roughly 20% or more) of players who purchase *no* edges, and another large group who purchase very little compared to the average. Thus the vast majority of the cost of building the networks was undertaken by only a small fraction of the population. This variance in behavior is what we believe generated the aforementioned variance in C_B values.

For low and medium edge costs, fully 50% of the population contributed less than 10% of the total edge expenditures; at high edge costs, where no player could afford to purchase more than 3 edges and thus variability of expenditure should be reduced, the free riding 50% still purchase less than 20% of the edges. Furthermore, free riding was an economically beneficial policy: at the level of individual human subjects, the correlation between the number of edges purchased in all Session 1 experiments and their total payoff is -0.72 ($P < 0.001$). The primary builders of the networks were thus apparently not financially favored by their resulting positions in it. Networks tended to be built rapidly, with the vast majority of edge purchases coming in the first half of the allowed time.

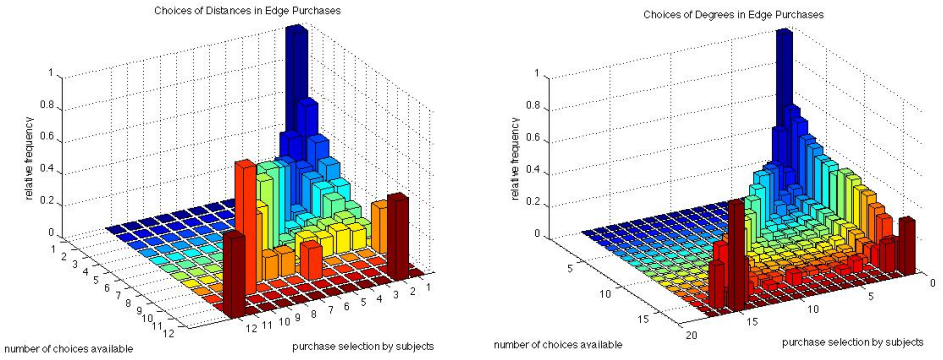


Fig. 6. Frequencies of edge-purchasing decisions with respect to non-neighbor vertex distances (left panel) and degrees (right panel). For each edge purchase, the left axis represents how many distinct choices the purchaser had, and the right axis represents which of these ranked options they selected. The vertical axis then shows the relative frequency they made each ranked choice. Thus the diagonals indicate cases where they purchased an edge to the most distant, or highest-degree, non-neighbor, respectively. We thus see that while there is a tendency to purchase edges to the most distant and highest degree vertices, there is also considerable mass at low and intermediate distances and degrees. Color is for visualization clarity only.

So far our emphasis has been on the structural properties of the built networks and the distribution of expenditures; we next examine the criteria subjects seemed to use in edge-purchasing decisions, within the constraints of the degree and distance information they were provided about current non-neighbors. Normalization is an issue here since (for instance) what constitutes a relatively "high degree" vertex is different near the start of an experiment than towards the end. Instead we can simply ask, at each moment an edge purchase was made, how many choices the purchaser had in each dimension, and which one they made: that is, how many different degree values, and how many different distance values, were populated by at least one non-neighbor on the edge-purchasing GUI grid. The results are summarized in Figure 6, and they show that while subjects most often chose to buy edges to vertices with the highest available distance, or the largest available degree, they also frequently chose deliberately low values in both dimensions as well, and there is significant mass on intermediate values as well. Thus purchasing behavior is not easily consistent with standard generative models such as Erdös-Renyi (which would induce uniform distributions in both dimensions) or preferential attachment (which would not generate the observed tendency to connect to low-degree vertices). Despite the simplicity of the informational interface, subject purchasing behavior does not easily fall into simple models.

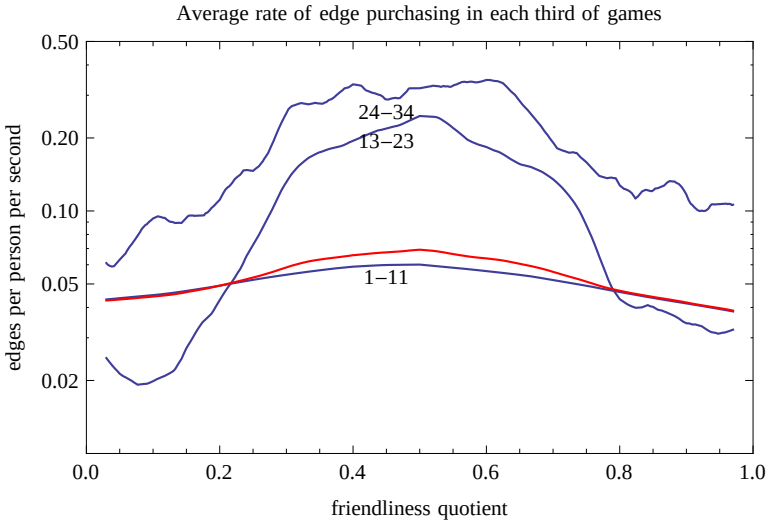


Fig. 7. Edge purchasing rate (edges purchased per subject per second) as a function of the friendliness quotient (see text) of the purchaser at the moment of purchase. The red curve shows the aggregate across all Session 1 and 2 purchases, and the peak near 0.5 is 60 or 70% higher than at the edges. If we condition on the purchaser also having a current degree in some range (blue curves annotated by degree range), we see that this tendency to purchase more when there is local indecision and conflict becomes greatly pronounced at higher degrees. The two higher degree curves are 5 to 10 times higher in the middle than at the sides.

While the preceding analysis examines how subjects used degree and distance in edge purchasing decisions, it is also of interest to investigate how such decisions were influenced by the local state of play — in particular, whether most neighbors were playing the subject's preferred (higher payoff) color or not. Figure 7 shows the rate at which subjects purchased edges in Session 1 and 2 experiments as a function of the "friendliness quotient" of their neighborhood at the moment of purchase. The friendliness quotient is the fraction of current neighbors who are playing the purchaser's higher-payoff color. We see that there is a marked increase in the proclivity of a player to buy an edge if she finds herself with an approximately equal number of friends and enemies (friendliness quotient 0.5). In situations where her neighbors are largely colored the same (whether of the higher or lower payoff color), she refrains from buying. This tendency becomes even more pronounced if we condition on just those purchases made by players whose current degree is higher. It may be that this tendency to buy more edges when there is a large amount of local disagreement (high degree, friendliness quotient close to 0.5) only worsened the indecision for everyone, thus leading to poor convergence performance.

5. CONCLUDING REMARKS

The results presented here have shown that human subjects, given the opportunity and incentives to collectively build a network in service of a competitive coordination task, did so poorly, creating networks inherently difficult for the task. This occurred despite an edge purchasing interface that permitted, in principle, the creation of networks known from earlier experiments to be much easier, such as random or preferential attachment networks. Our findings are in contrast to some case studies and

theories about the abilities of human populations to effectively self-organize in service of collective goals. This contrast clearly deserves further scrutiny and controlled study.

REFERENCES

ALBERS, S., EILTS, S., EVEN-DAR, E., MANSOUR, Y., AND RODDITY, L. 2006. On Nash equilibria for a network creation game. In *Symposium on Discrete Algorithms (SODA)*.

BARABASI, L. AND ALBERT, R. 1999. Emergence of scaling in random networks. *Science 286*, 509–512.

BOLLABAS, B. 2001. *Random Graphs*. Cambridge University Press.

BORGS, C., CHAYES, J., DING, J., AND LUCIER, B. 2011. The hitchhiker's guide to affiliation networks: A game-theoretic approach. In *Innovations in Theoretical Computer Science (ITCS)*.

BRAUTBAR, M. AND KEARNS, M. 2011. A clustering coefficient network formation game. In *Symposium on Algorithmic Game Theory (SAGT)*.

BURNS, T. AND STALKER, G. 1994. *The Management of Innovation*. Oxford University Press.

CHAKRABORTY, T., JUDD, J. S., KEARNS, M., AND TAN, J. 2010. A behavioral study of bargaining in social networks. In *ACM Conference on Electronic Commerce*. 243–252.

EVEN-DAR, E. AND KEARNS, M. 2006. A small world threshold for economic network formation. In *Neural Information Processing Systems (NIPS)*.

EVEN-DAR, E., KEARNS, M., AND SURI, S. 2007. A network formation game for bipartite exchange economies. In *Symposium on Discrete Algorithms (SODA)*.

FABRIKANT, A., LUTHRA, A., MANEVA, E., PAPADIMITRIOU, C., AND SHENKER, S. 2003. On a network creation game. In *Principles of Distributed Computing (PODC)*.

JACKSON, M. O. 2005. A survey of models of network formation: stability and efficiency. In *Group Formation in Economics: Networks, Clubs and Coalitions*. Cambridge University Press.

JACKSON, M. O. 2010. *Social and Economic Networks*. Princeton University Press.

JUDD, S., KEARNS, M., AND VOROBEYCHIK, Y. 2010. Behavioral dynamics and influence in networked coloring and consensus. *Proceedings of the National Academy of Sciences 107,* 34, 14978–14982.

KEARNS, M. AND JUDD, J. S. 2008. Behavioral experiments in networked trade. In *ACM Conference on Electronic Commerce*. 150–159.

KEARNS, M., JUDD, S., TAN, J., AND WORTMAN, J. 2009. Behavioral experiments in biased voting in networks. *Proceedings of the National Academy of Sciences 106,* 5, 1347–1352.

KEARNS, M., JUDD, S., AND VOROBEYCHIK, Y. 2011. Behavioral conflict and fairness in social networks. In *Workshop on Internet and Network Economics*.

KEARNS, M., SURI, S., AND MONTFORT, N. 2006. An experimental study of the coloring problem on human subject networks. *Science 313*, 824–827.

KLEINBERG, J. 2000. Navigation in a small world. *Nature 406*, 845.

NISHIGUCHI, T. AND BEAUDET, A. 2000. Fractal design: Self-organizing links in supply chain management. In *Knowledge Creation: A Source of Value*, G. von Krogh et. al., Ed. St. Martin's Press, 199–230.

TARDOS, E. AND WEXLER, T. 2007. Network formation games and the potential function method. In *Algorithmic Game Theory*. Cambridge University Press, 487–513.

WATTS, D. AND STROGATZ, S. 1998. Collective dynamics of small-world networks. *Nature 393*, 409–410.

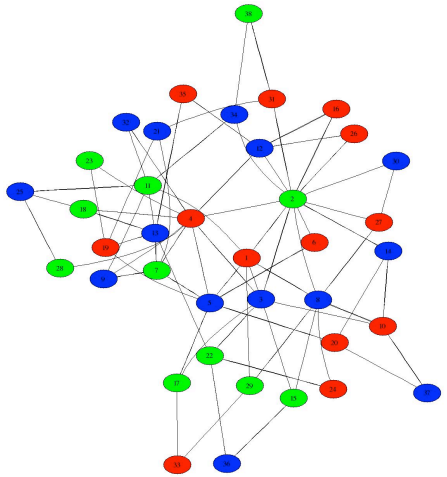
Trading in Networks: I. Model

Prof. Michael Kearns

Networked Life

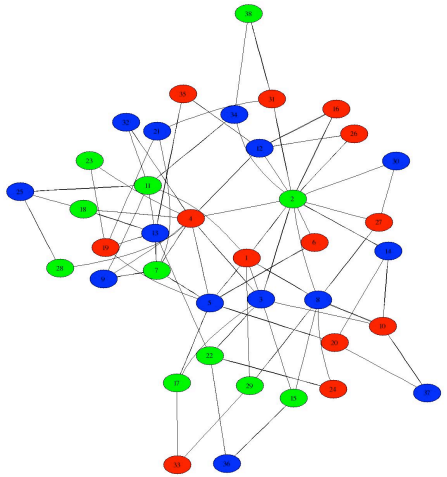
NETS 112

Fall 2014



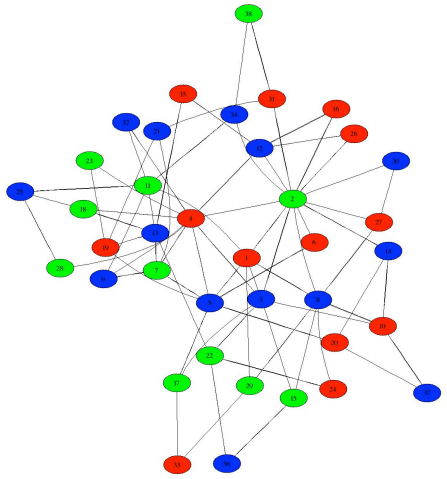
Roadmap

- Networked trading motivation
- A simple model and its equilibrium
- A detailed example



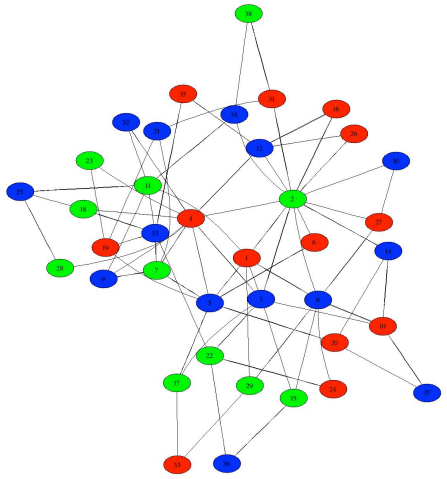
Networked Games vs. Trading

- Models and experiments so far (coloring, consensus, biased voting):
  - simple coordination games
  - extremely simple actions (pick a color)
  - "trivial" equilibrium theories ("good" equilibrium or "trapped" players)
  - no equilibrium predictions about network structure and individual wealth
- Networked trading:
  - a "financial" game
  - complex action space (set of trades with neighbors)
  - nontrivial equilibrium theory
  - detailed predictions about network structure and individual wealth



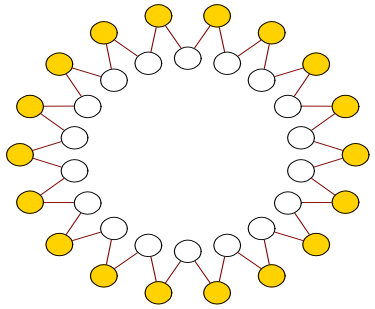
Networked Trading: Motivation

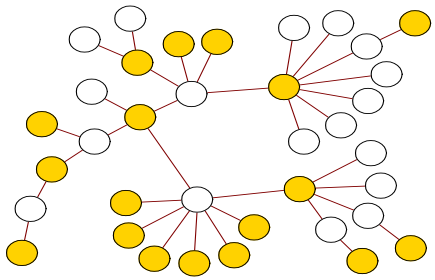
- Settings where there are restrictions on who can trade with whom
- International trade: restrictions, embargos and boycotts
- Financial markets: some transactions are forbidden
  - e.g. trades between brokerage and proprietary trading in investment banks
- Geographic constraints: must find a local housecleaning service
- Natural to model by a network:
  - vertices representing trading parties
  - presence of edge between u and v: trading permitted between parties
  - absence of edge: trading forbidden



A Simple Model of Networked Trading

- Imagine a world with only two goods or commodities for trading
  - let's call them Milk and Wheat
- Two types of traders:
  - Milk traders: start game with 1 unit (fully divisible) of Milk, but only value Wheat
  - Wheat traders: start game with 1 unit of Wheat, but only value Milk
  - trader's payoff = amount of the "other" good they obtain through trades
  - "mutual interest in trade"
  - equal number of each type → same total amount of Milk and Wheat
- Only consider *bipartite* networks:
  - all edges connect a Milk trader to a Wheat trader
  - can only trade with your network neighbors!
  - all trades are irrevocable
  - no resale or arbitrage allowed

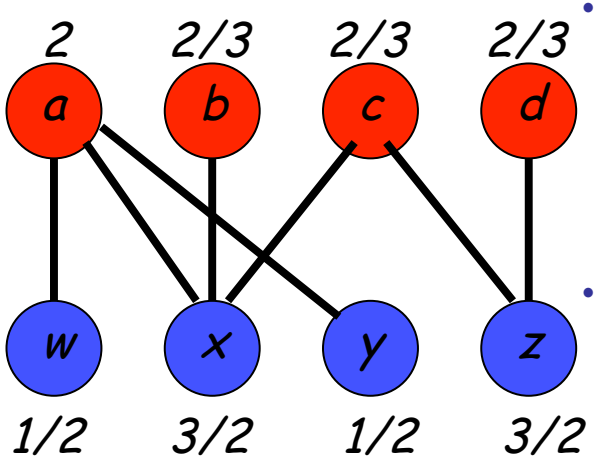




Equilibrium Concept

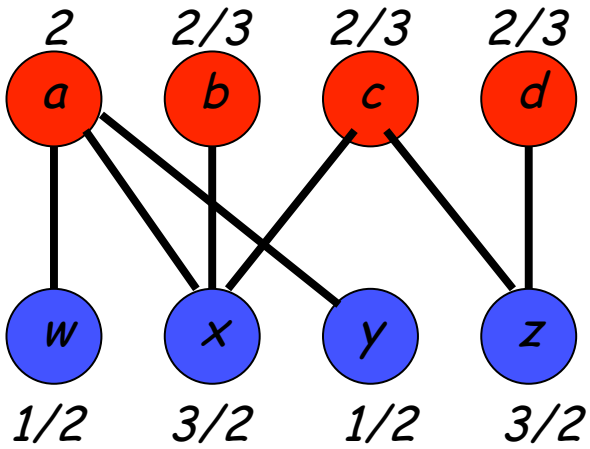
- Imagine we assigned a price or exchange rate to each vertex/trader
  - e.g. "I offer my 1 unit of Milk for 1.7 units of Wheat"
  - e.g. "I offer my 1 unit of Wheat for 0.8 units of Milk"
  - note: "market" sets the prices, not traders ("invisible hand")
  - unlike a traditional game --- traders just react to prices
- Equilibrium = set of prices + trades such that:
  - 1. market *clears:* everyone trades away their initial allocation
  - 2. rationality (best responses): a trader only trades with best prices in neighborhood
  - e.g. if a Milk trader's 4 neighbors offer 0.5, 1.0, 1.5, 1.5 units Wheat, they can trade only with those offering 1.5
  - note: set of trades must ensure supply = demand at every vertex
- Simplest example: complete bipartite network
  - every pair of Milk and Wheat traders connected by an edge
  - equilibrium prices: everyone offers their initial 1 unit for 1 unit of the other good
  - equilibrium trades: pair each trader with a unique partner of other type
  - market clears: everyone engages in 1-for-1 trade with their partner
  - rationality: all prices are equal, so everyone trading with best neighborhood prices

A More Complex Example

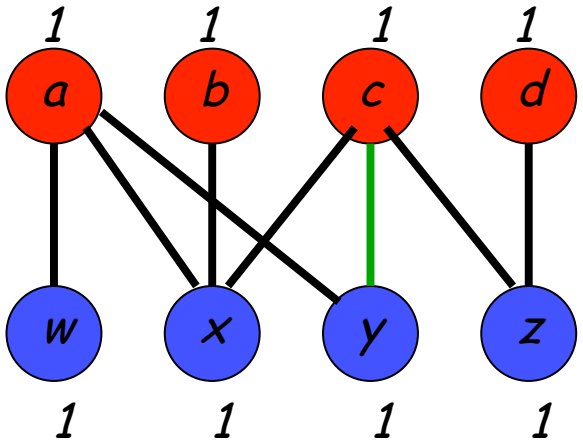


- equilibrium prices as shown (amount of the other good demanded)
- equilibrium trades:
  - a: sends $\frac{1}{2}$ unit each to w and y, gets 1 from each
  - b: sends 1 unit to x, gets 2/3 from x
  - c: sends $\frac{1}{2}$ unit each to x and z, gets 1/3 from each
  - d: sends 1 unit to z, gets 2/3 from z
- equilibrium check, blue side:
  - w: traded with a, sent 1 unit
  - x: traded with b and c, sent 1 unit
  - y: traded with a, sent 1 unit
  - z: traded with c and d, sent 1 unit

Remarks



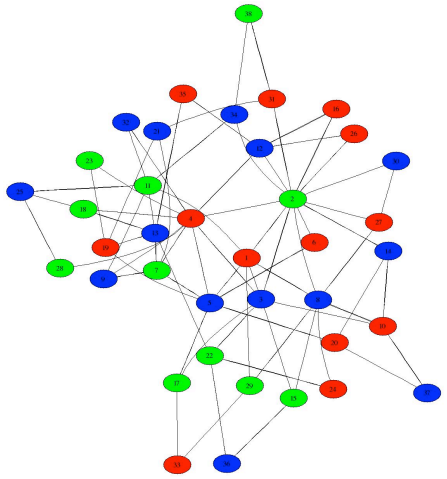
- How did I figure this out? Not easy in general
- Some edges unused by equilibrium
- Trader wealth = equilibrium price at their vertex
- If two traders trade, their wealths are reciprocal (w and 1/w)
- Equilibrium *prices (wealths)* are always unique
- Network structure led to *variation* in wealth



- Suppose we add the single green edge
- Now equilibrium has no wealth variation!

Summary

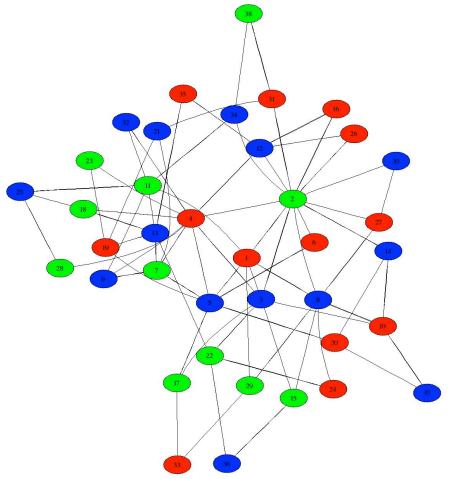
- (Relatively) simple networked trading model
- Equilibrium = prices + trades such that market clears, traders rational
- Some networks don't have wealth variation at equilibrium, some do
- Next: What is the general relationship between structure and prices?



Trading in Networks:
II. Network Structure and Equilibrium

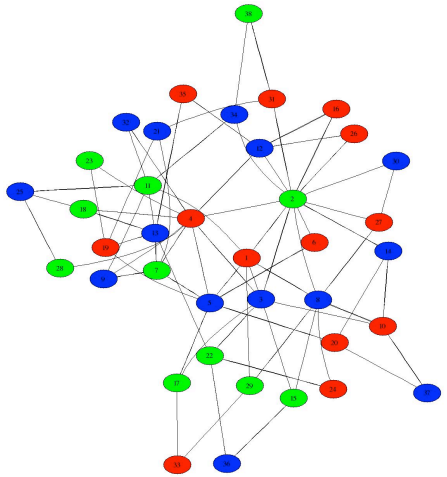
Networked Life

Prof. Michael Kearns



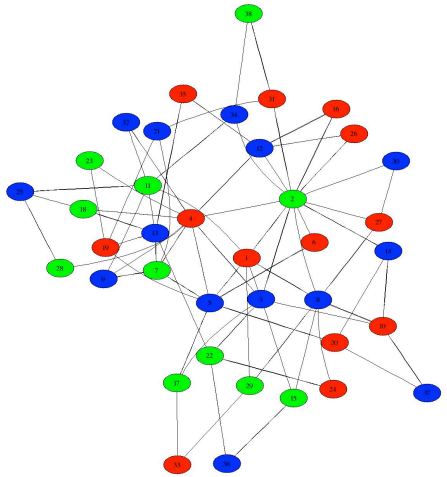
Roadmap

- Perfect matchings and equilibrium equality
- Characterizing wealth inequality at equilibrium
- Economic fairness of Erdös-Renyi and Preferential Attachment

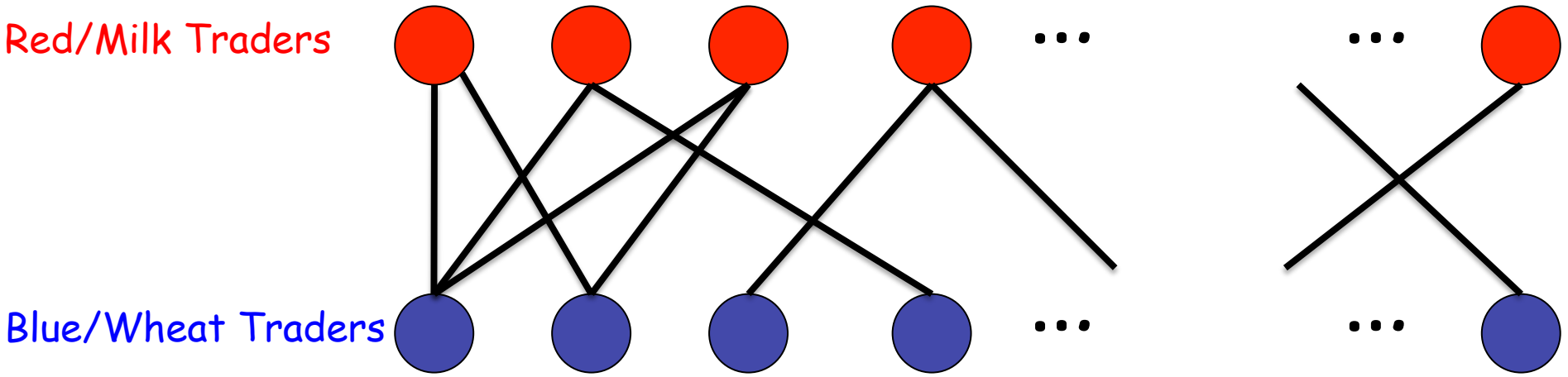


Trading Model Review

- Bipartite network, equal number of Milk and Wheat traders
- Each type values only the other good
- Equilibrium = prices + trades such that market clears, traders rational

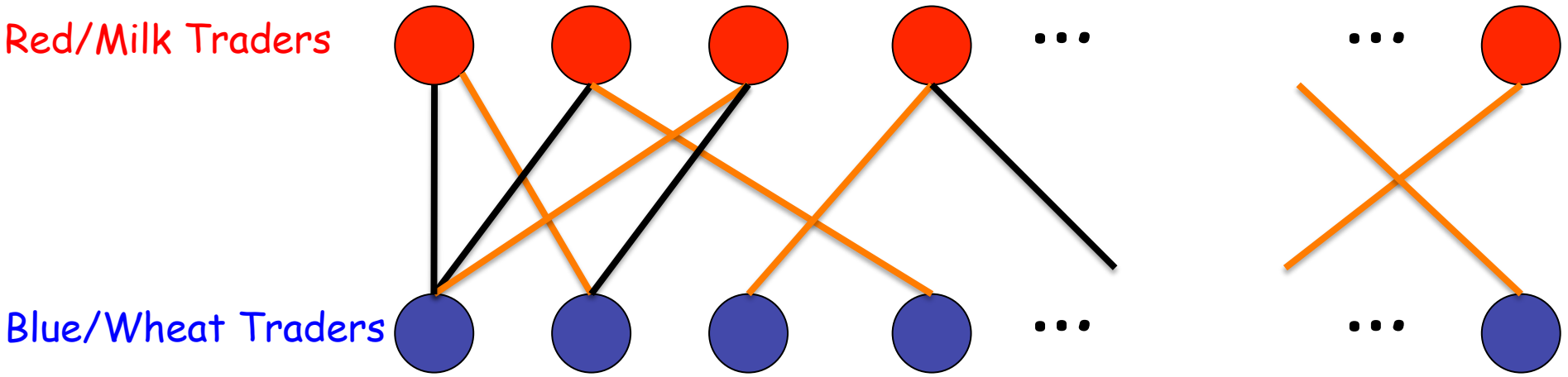


Perfect Matchings



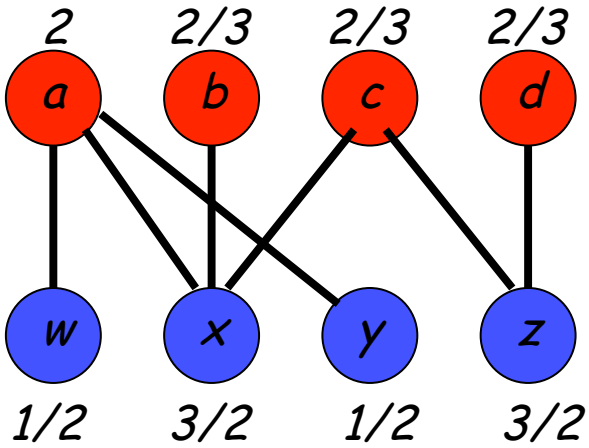
- A pairing of reds and blues so everyone has *exactly one partner*
- So really a subset of the edges with each vertex in exactly one edge
- Some networks may have many different perfect matchings
- Some networks may have no perfect matchings

Perfect Matchings

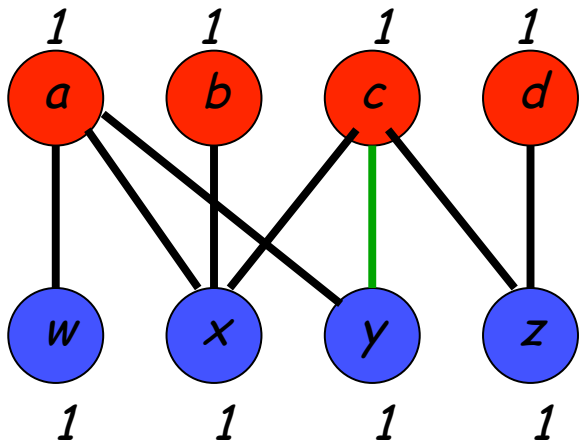


- A pairing of reds and blues so everyone has *exactly one partner*
- So really a subset of the edges with each vertex in exactly one edge
- Some networks may have many different perfect matchings
- Some networks may have no perfect matchings

Examples



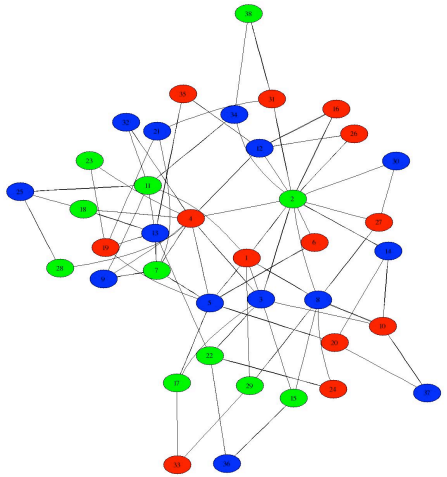
Has no perfect matching



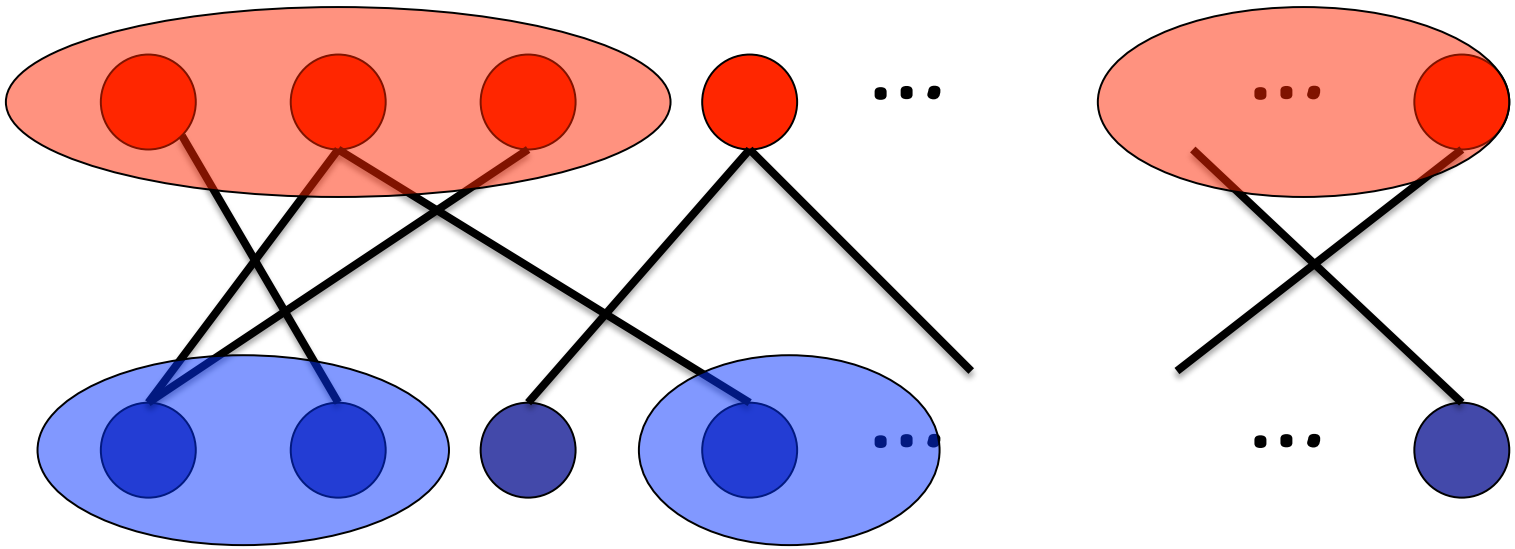
Has a perfect matching

Perfect Matchings and Equality

- Theorem: There will be *no wealth variation* at equilibrium (all exchange rates = 1) if and only if the bipartite trading network contains a perfect matching.
- Characterizes sufficient "trading opportunities" for fairness
- What if there is no perfect matching?



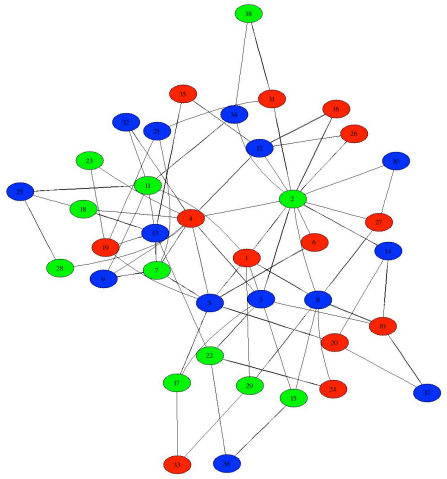
Neighbor Sets



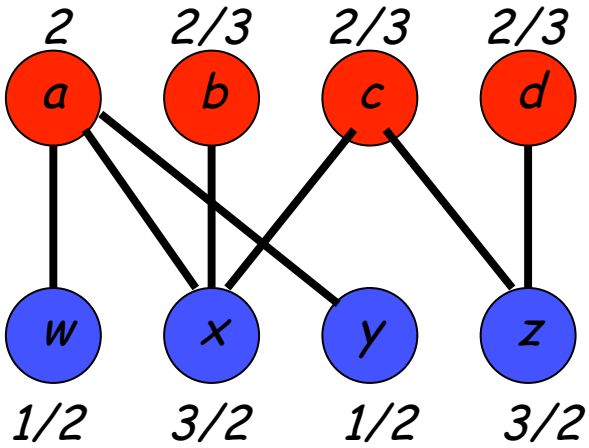
- Let S be any set of traders on one side
- Let N(S) be the set of traders on the other side connected to any trader in S; these are the only trading partners for S collectively
- Intuition: if N(S) is much smaller than S, S may be in trouble
- S are "captives" of N(S)
- Note: If there is a perfect matching, N(S) always *at least as large* as S

Characterizing Inequality

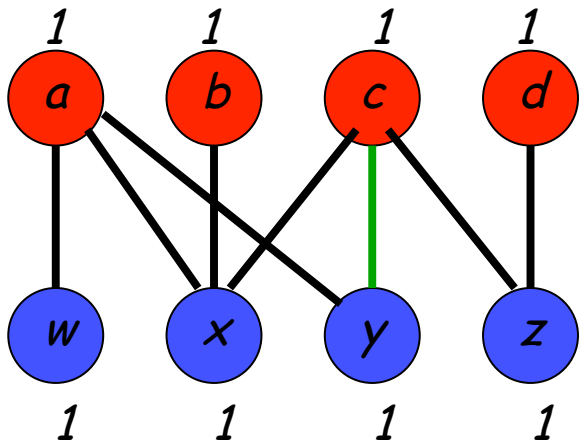
- For any set S, let v(S) denote the ratio (size of S)/(size of N(S))
- Theorem: If there is a set S such that v(S) > 1, then at equilibrium the traders in S will have wealth at most 1/v(S), and the traders in N(S) will have wealth at least v(S).
- Example: v(S) = 10/3 → S gets at most 3/10, N(S) at least 10/3
- Greatest inequality: find S maximizing v(S)
- Can iterate to find all equilibrium wealths
- Corollary: adding edges can only *reduce* inequality
- Network structure completely determines equilibrium wealths
- Note: trader/vertex degree not directly related to equilibrium wealth



Examples Revisited



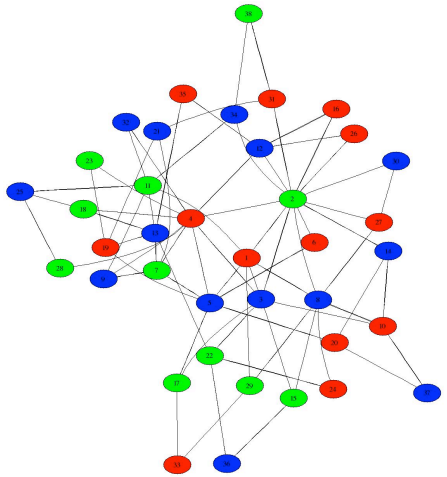
Has no perfect matching



Has a perfect matching

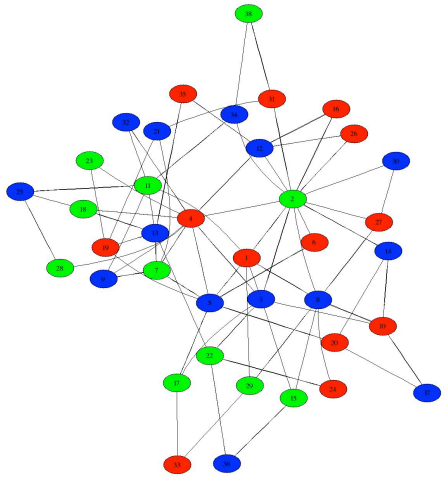
Inequality in Formation Models

- Bipartite version of Erdös-Renyi: even at low edge density, very likely to have a perfect matching → *no wealth variation* at equilibrium
- Bipartite version of Preferential Attachment: wealth variation will *grow rapidly* with population size
- Erdös-Renyi generates economically "fairer" networks



Summary

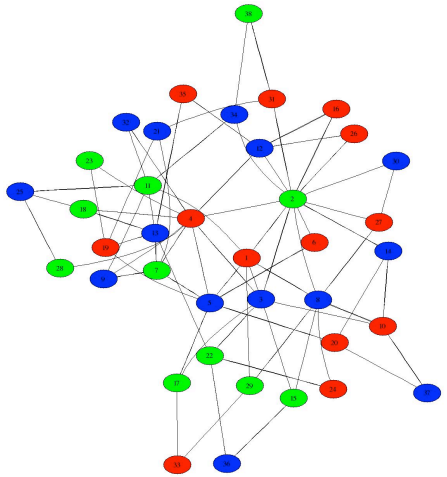
- Ratios v(S) completely characterize equilibrium
- Determined entirely by network structure
- More subtle and global than trader degrees
- Next: comparing equilibrium predictions with human behavior



Trading in Networks:
III. Behavioral Experiments

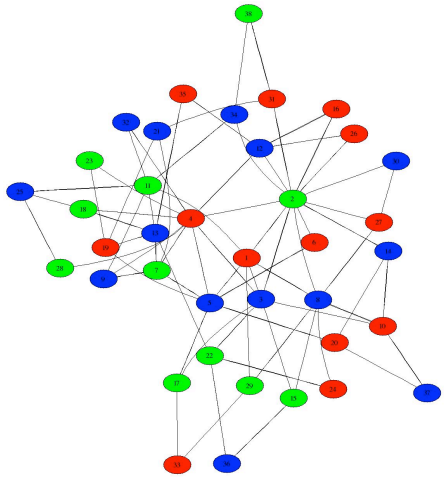
Networked Life

Prof. Michael Kearns



Roadmap

- Experimental framework and trading mechanism/interface
- Networks used in the experiments
- Visualization of actual experiments
- Results and comparison to equilibrium theory predictions

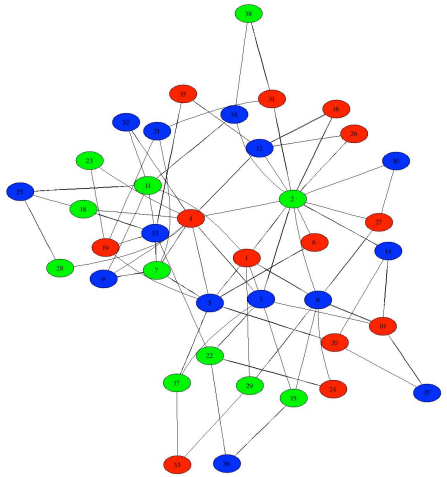


Equilibrium Theory Review

- Equilibrium prices/wealths entirely determined by network structure
- Largest/smallest wealths determined by largest ratios:

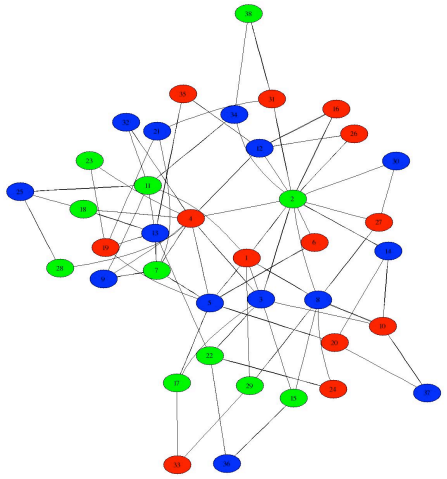
v(S) = (size of S)/(size of N(S)) N(S) "winners", S "losers"

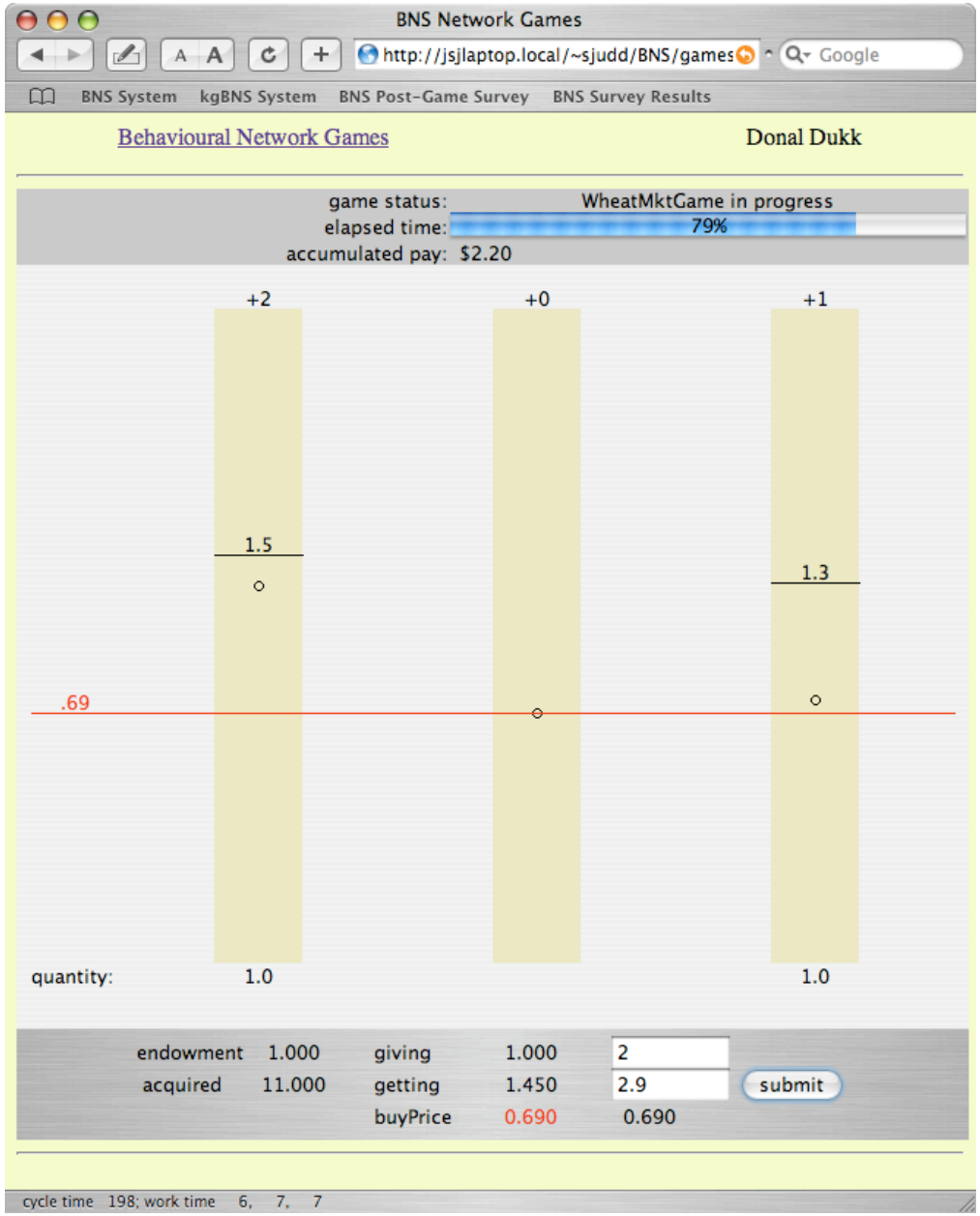
- Network has a perfect matching: all wealths = 1

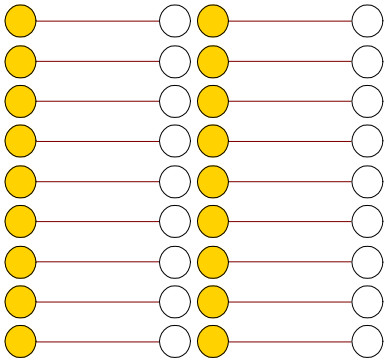


Experimental Framework

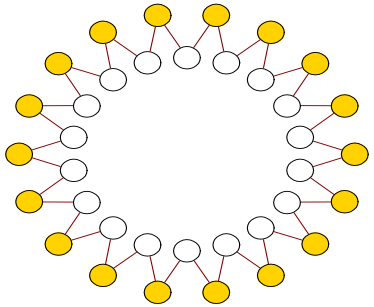
- Same framework as coloring, consensus and biased voting experiments
- 36 simultaneous human subjects in lab of networked workstations
- In each experiment, subjects play our trading model on varying networks
- In equilibrium theory, prices are magically *given* ("invisible hand")
- In experiments, need to provide a mechanism for price *discovery*
- Experiments used simple *limit order* trading with neighbors
  - networked version of standard financial/equity market mechanism
- Each player starts with 10 fully divisible units of Milk or Wheat
  - payments proportional to the amount of the other good obtained



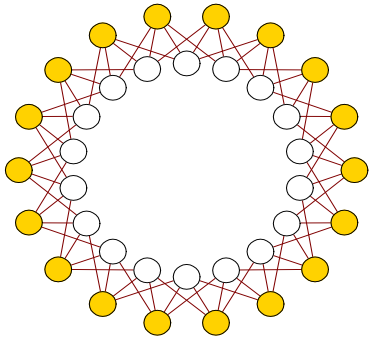

BNS Network Games
http://jsjlaptop.local/~sjudd/BNS/games
Google
BNS System
kgBNS System
BNS Post-Game Survey
BNS Survey Results
Behavioural Network Games
Donal Dukk
game status: WheatMktGame in progress
elapsed time: 79%
accumulated pay: $2.20
+2
+0
+1
1.5
1.3
.69
quantity:
1.0
1.0
endowment 1.000
acquired 11.000
giving 1.000
getting 1.450
buyPrice 0.690
2
2.9
0.690
submit
cycle time 198; work time 6, 7, 7



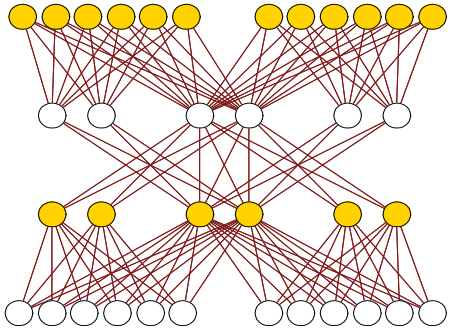
Pairs



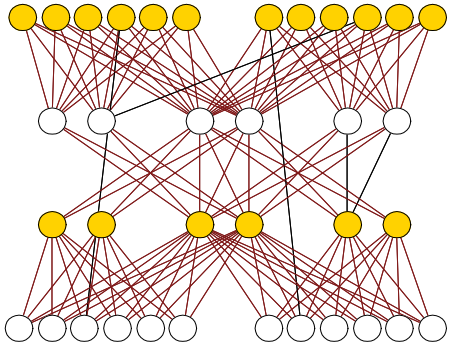
2-Cycle



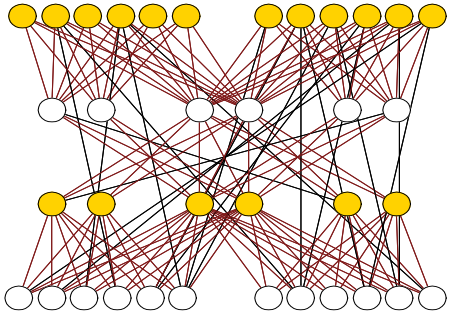
4-Cycle



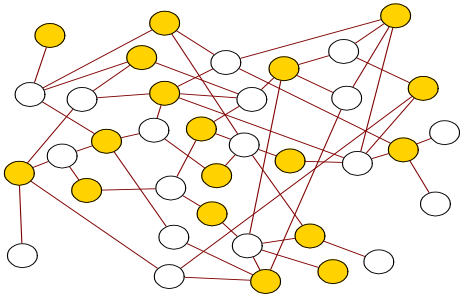
Clan



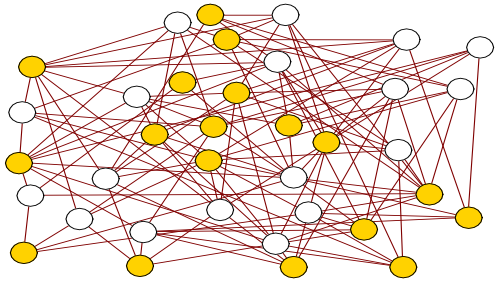
Clan + 5%



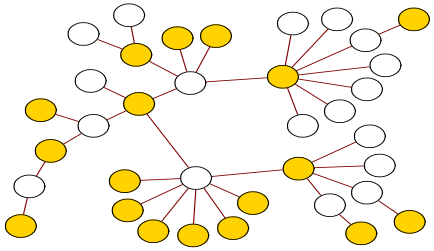
Clan + 10%



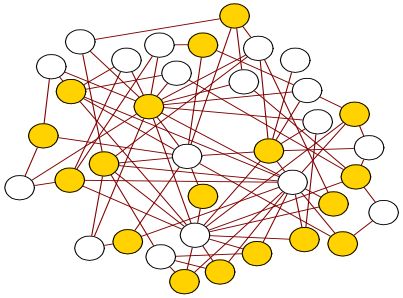
Erdos-Renyi, p=0.2



E-R, p=0.4



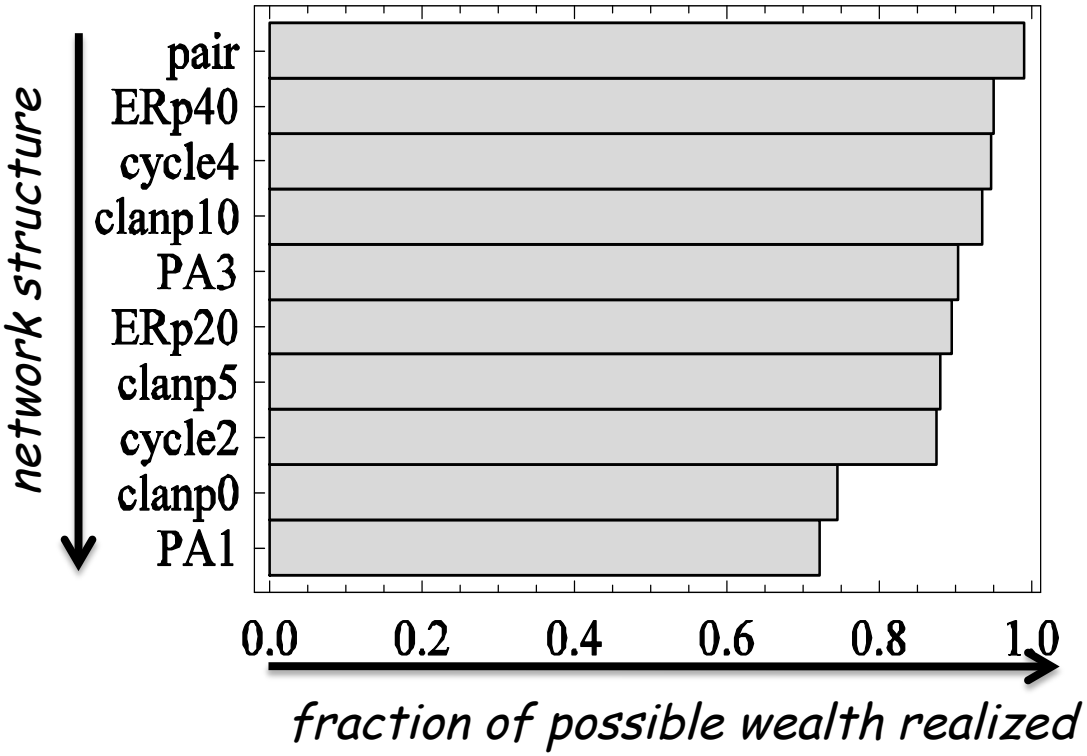
Pref. Att. Tree



Pref. Att. Dense

[movies]

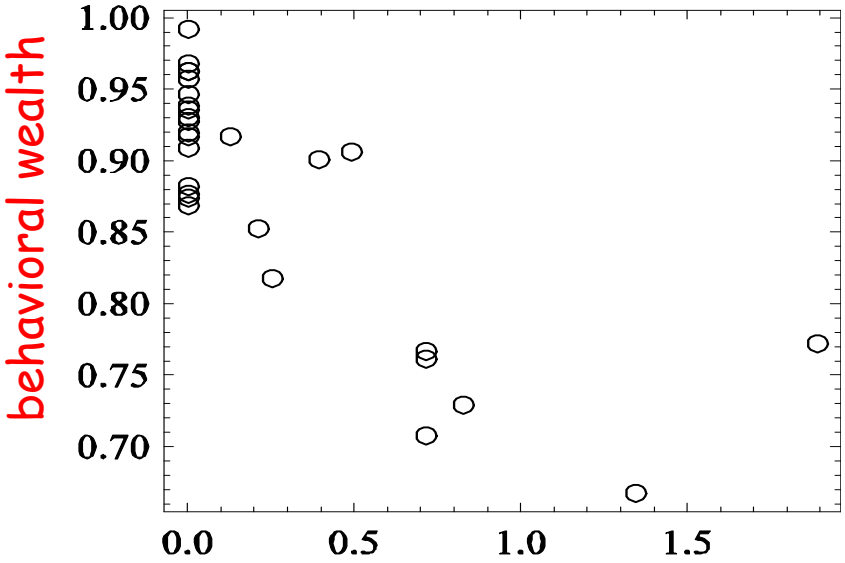
Collective Performance and Structure

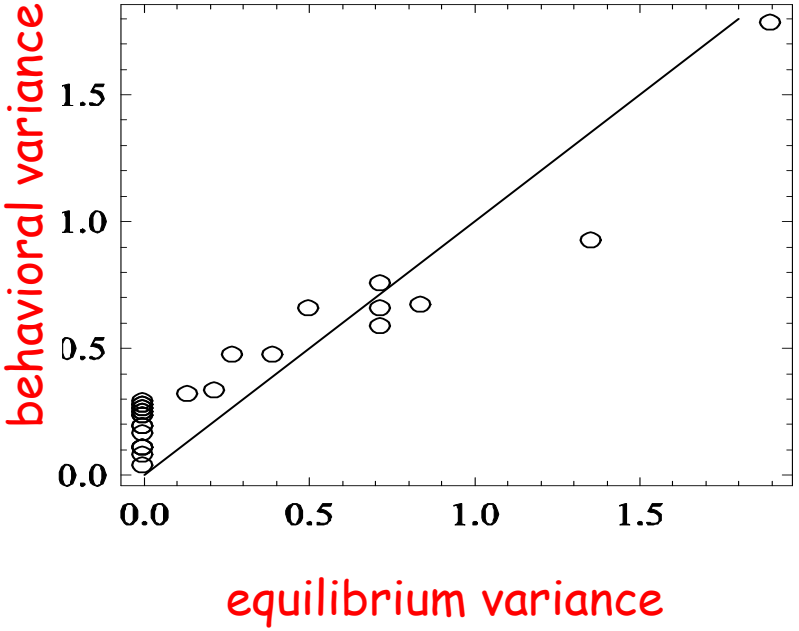


overall mean ~ 0.88

- overall behavioral performance is strong
- structure matters; many (but not all) pairs distinguished

Equilibrium vs. Behavior





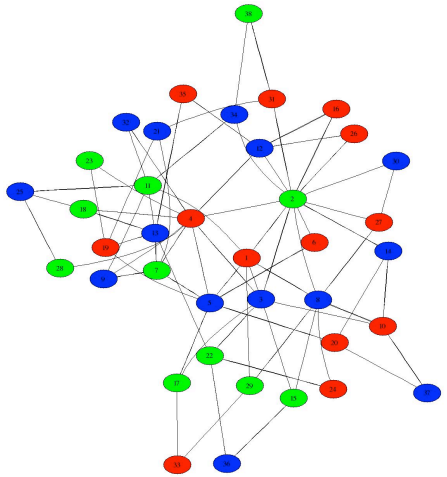
correlation ~ -0.8 ($p < 0.001$)

correlation ~ 0.96 ($p < 0.001$)

- greater equilibrium variation → behavioral performance degrades
- greater equilibrium variation → greater behavioral variation

Best Model for Behavioral Wealths?

- The equilibrium wealth predictions are better than:
  - degree distribution and other centrality/importance measures
  - uniform distribution
- Best behavioral prediction: 0.75(equilibrium prediction) + 0.25(uniform)
- "Networked inequality aversion" (recall Ultimatum Game)



Summary

- Trading model most sophisticated "rational dynamics" we've studied
- Has a detailed equilibrium theory based entirely on network structure
- Equilibrium theory matches human behavior pretty well