LECTURE 2:SOCIAL MEDIA AND NETWORKS

Prof. Pan Hui

CSIT 6000K: Social Networks and Social Computing: A Data Science Perspective

Thursdays 07:30 PM - 10:20 PM

Social media

- Social media has transformed society
  - Reduced barriers to communication
  - Democratized content publication

- As a computer scientist…
  - Tend to ignore users
  - Social media makes users a part of the system

- Important to understand interactions
  - Within the system (traditional CS)
  - Between users and system (HCI)
  - Among users themselves (sociology)

What is social media?

- Systems with user interaction as critical component

- Online communities
  - Facebook, MySpace, YouTube

- Communication systems
  - Skype, WhatsApp, WeChat, Line

- Social news media
  - Blogs, iReport, Instgram

- Online worlds
  - World of WarCraft, Second Life

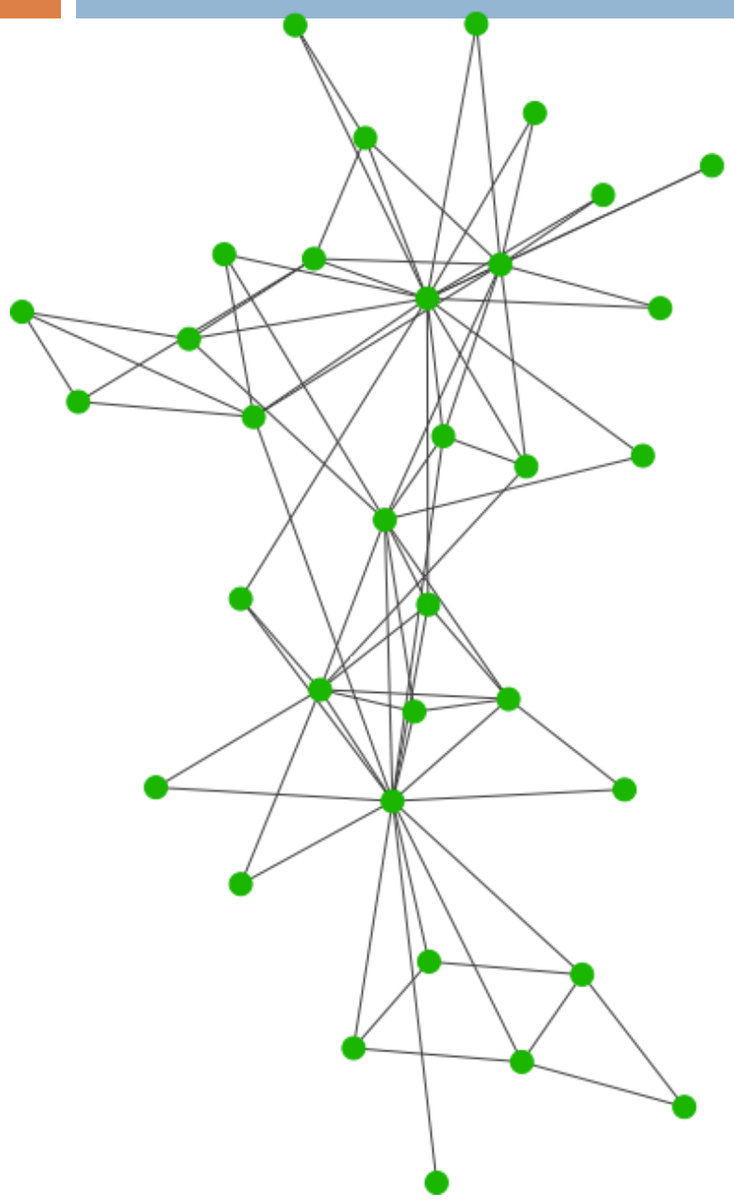
Why is social media interesting?

- Two reasons (to me):

- 1. Observe social interaction at scale
  - Social media based user interactions
  - Scale not possible before

- 2. Relate information and people
  - Online social networks now content-sharing systems
  - Can attach reputation of users to content

1. Observe social interactions



Anyone recognize this network?

- Zachary's Karate Club

- Collecting it involved massive field work
  - Manually observe people
  - Trace interactions for two years (!)
  - Will discuss more later

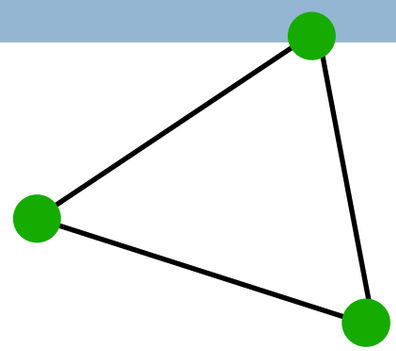
- Limit in scalability of this approach
  - Biases from interviewing
  - Time spent

Opportunity: Large-scale data

- An opportunity to scale up observations
  - "Field work" required may be reduced
- Social media sites have complete history record
  - Interactions, discussions, friendship creation (and deletion), …
  - Entire evolution of a group of users
- At incredible detail
  - 74% of Facebook users login at least once a day
  - 20B people-minutes spent on Facebook per day
  - Every interaction recorded

The curse of scale

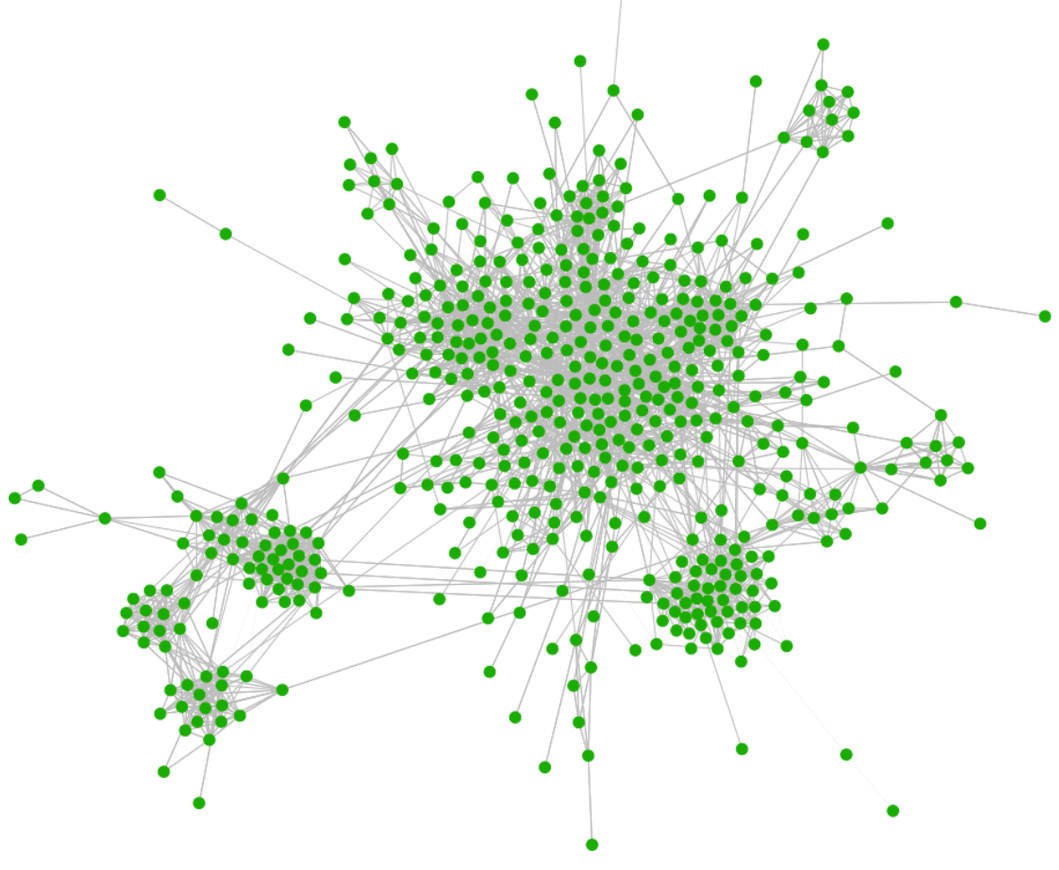
- Scale is both a blessing and a curse



- Blessing
  - Confidence in results
  - Certain effects only seen at scale

- Curse
  - Miss many local interactions
  - Links “mean” less
  - Comparing networks hard

- Important to keep limitations in mind

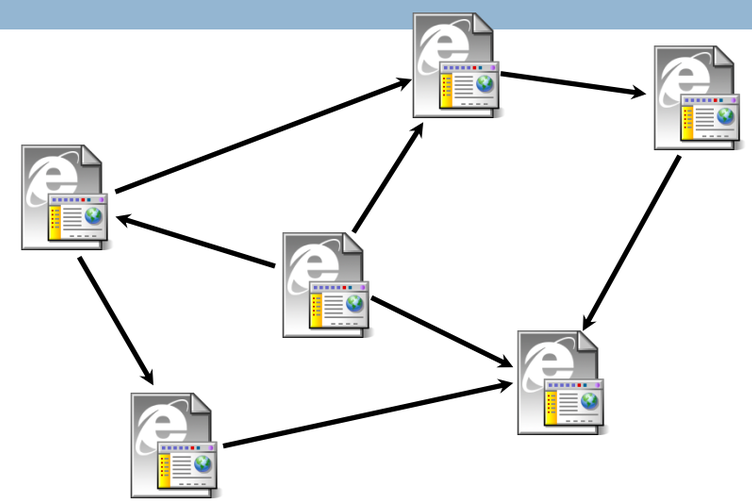


2. Relate information to people

- Popular way to connect and share content
  - Photos, videos, blogs, profiles, news, status…
  - Twitter (330 M), Facebook (2.7 B)
- Growing exponentially
- Incredible amounts of content being shared
  - Facebook (350 M photos/day)
  - YouTube (500 hours of video/min)

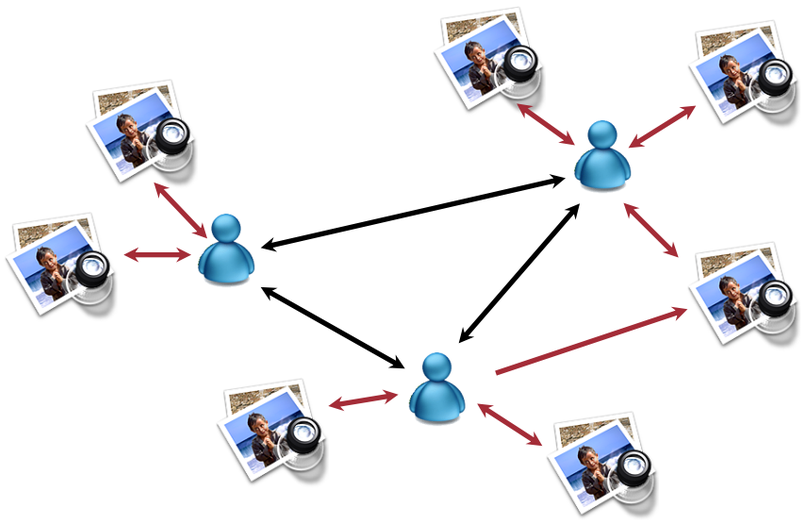
A new way of organizing information

- Web organized with content-content links
  - Link structure exploited (e.g., PageRank)



Web

- Social media organized using
  - User-user links (social network)
  - User-content links (favorites, etc)

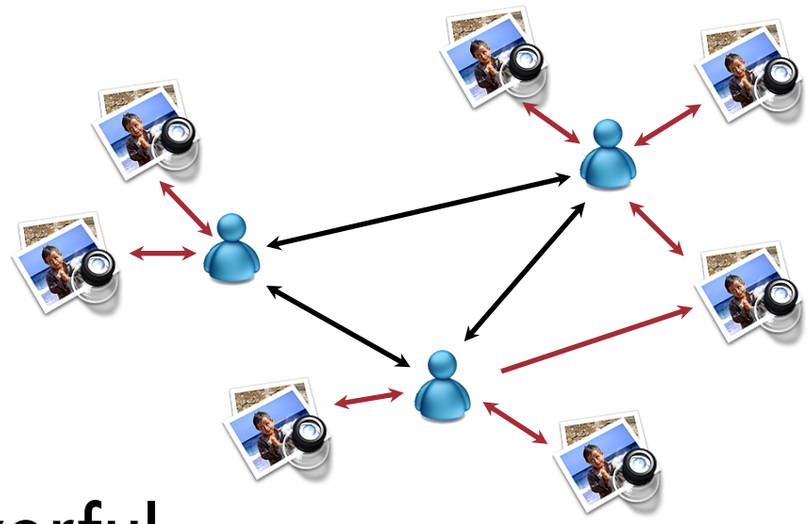


Social networks

- New platform for information sharing

Relates information to people

- Today, social network used to structure information

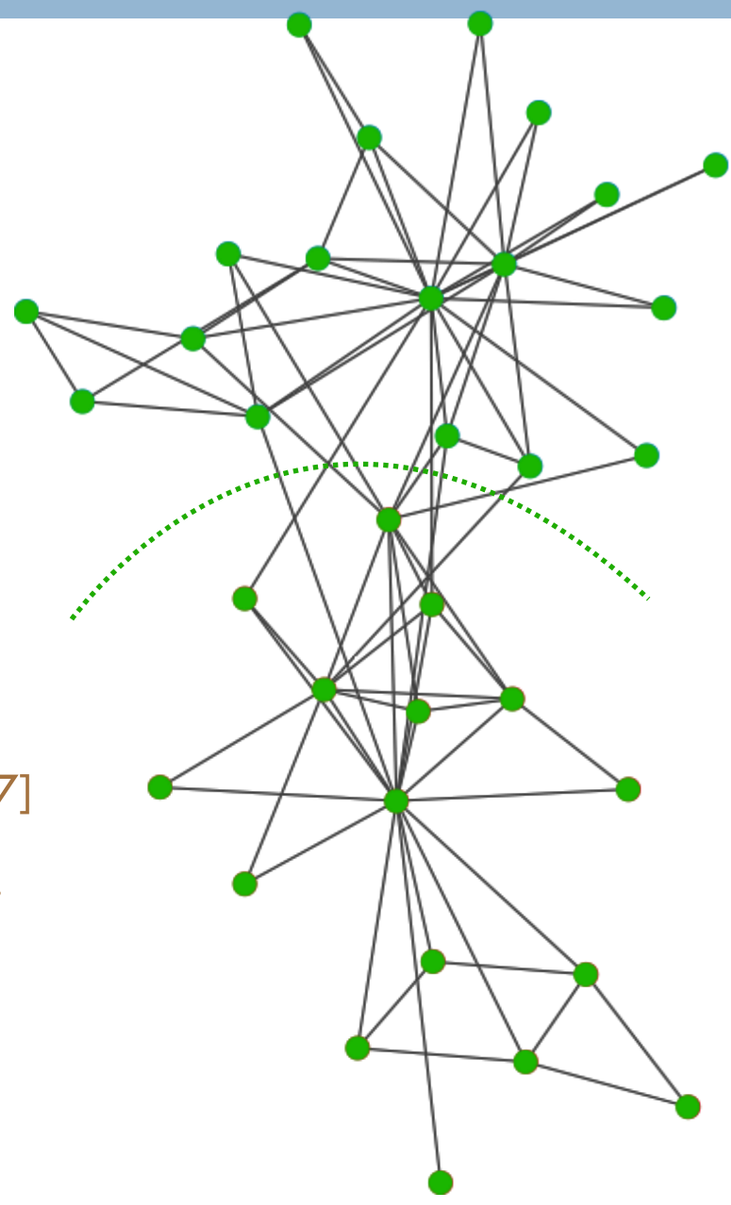


- Can we extract other information?
  - Combination of who and what very powerful

- Social network connects content with
  - (Multiple) user's reputation
  - Community the user is part of

But why study networks?

- Does network science make sense for social media?
  - Why not study interactions directly?
- Natural fit with interactions
  - Users only interact with small subset of others
- Degrees of influence beyond friends
  - Obesity [Fowler and Christakis, NE J. Med. 2007]
  - Altruism [Fowler et al., Econ. Let. 2009]
- Example: Zachary's Karate Club
  - Can predict behavior with network view
    - [Zachary, J. Anth. Res. 1972]



What sort of questions are we asking?

- Already know lots about networks
  - Scale-free [Barabasi and Albert, 1999], High clustering [Watts and Strogatz, 1998], Navigable [Adamic and Adar, 2003] [Liben-Nowell 2005], Hubs and authorities [Page and Brin, 1998] [Kleinberg, 1999], Dense core [Mislove et al. 2007]
- And have lots of models
  - Preferential Attachment [Barabasi and Albert, Nature 1999], Small-world [Watts and Strogatz, 1998], Copying [Kleinberg et al., 1999], Congestion [Mihail et al., 2003], Bowtie [Broder et al., 2000], Jellyfish [Tauro et al., 2001]
- Thus, going to focus on social aspects
  - Why do they look the way they do?
  - What can this tell us?

Outline

- Two parts:

- 1 Primer on social sciences

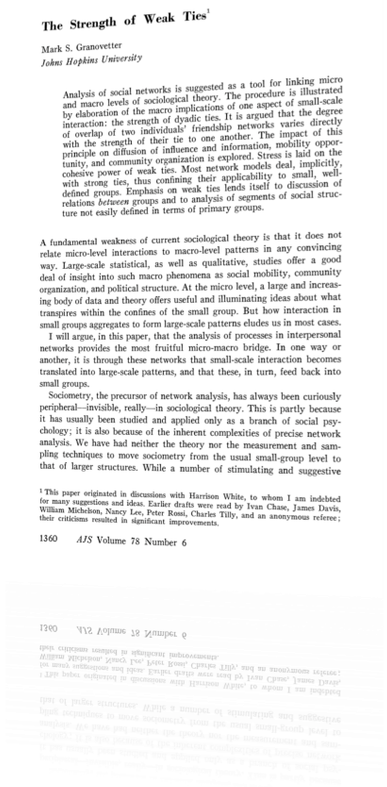
- 2 Leveraging social media

Goals

- Provide an overview of research on social media and networks
- Get you excited about this research area
- Give pointers to further reading
  - Papers cited throughout talk
- Spark discussion
  - Interrupt and ask questions!

PRIMER ON SOCIAL SCIENCES

Three classic papers



The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

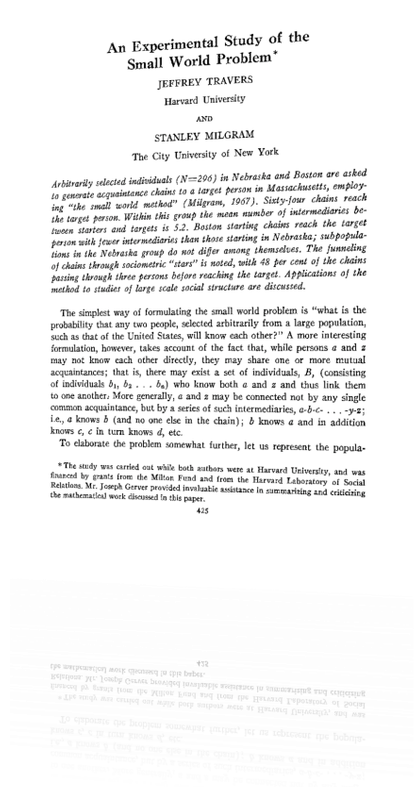
Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases.

I will argue, in this paper, that the analysis of processes in interpersonal networks provides the most fruitful micro-macro bridge. In one way or another, it is through these networks that small-scale interaction becomes translated into large-scale patterns, and that these, in turn, feed back into small groups.

Sociometry, the precursor of network analysis, has always been curiously peripheral—invisible, really—in sociological theory. This is partly because it has usually been studied and applied only as a branch of social psychology; it is also because of the inherent complexities of precise network analysis. We have had neither the theory nor the measurement and sampling techniques to move sociometry from the usual small-group level to that of larger structures. While a number of stimulating and suggestive

[1] This paper originated in discussions with Harrison White, to whom I am indebted for many suggestions and ideas. Earlier drafts were read by Ivan Chase, James Davis, William Michelson, Nancy Lee, Peter Rossi, Charles Tilly, and an anonymous referee; their criticisms resulted in significant improvements.



An Experimental Study of the Small World Problem*

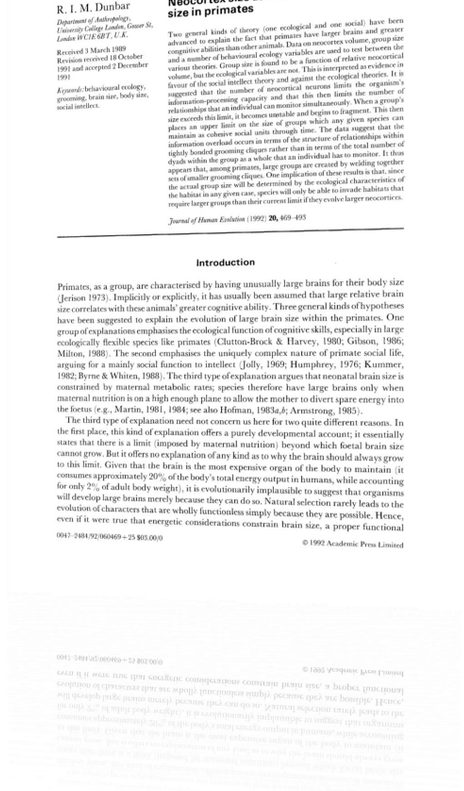
JEFFREY TRAVERS
Harvard University
AND
STANLEY MILGRAM
The City University of New York

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_i (consisting of individuals $b_1, b_2 \ldots b_n$) 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.



R. I. M. Dunbar
Department of Anthropology, University College London, Gower St, London WC1E 6BT, U.K.

Received 3 March 1989
Revision received 18 October 1991 and accepted 2 December 1991

Keywords: behavioural ecology, grooming, brain size, body size, social intellect.

Neocortex size as a constraint on group size in primates

Two general kinds of theory (one ecological and one social) have been advanced to explain the fact that primates have larger brains and greater cognitive abilities than other animals. Data on neocortex volume, group size and a number of behavioural ecology variables are used to test between the various theories. Group size is found to be a function of relative neocortical volume, but the ecological variables are not. This is interpreted as evidence in favour of the social intellect theory and against the ecological theories. It is suggested that the number of neocortical neurons limits the organism's information-processing capacity and that this then limits the number of relationships that an individual can monitor simultaneously. When a group's size exceeds this limit, it becomes unstable and begins to fragment. This then places an upper limit on the size of groups which any given species can maintain as cohesive social units through time. The data suggest that the information overload occurs in terms of the structure of relationships within tightly bonded grooming cliques rather than in terms of the total number of dyads within the group as a whole that an individual has to monitor. It thus appears that, among primates, large groups are created by wedding together sets of smaller grooming cliques. One implication of these results is that, since the actual group size will be determined by the ecological characteristics of the habitat in any given case, species will only be able to invade habitats that require larger groups than their current limit if they evolve larger neocortices.

Journal of Human Evolution (1992) **20**, 469–493

Introduction

Primates, as a group, are characterised by having unusually large brains for their body size (Jerison 1973). Implicitly or explicitly, it has usually been assumed that large relative brain size correlates with these animals' greater cognitive ability. Three general kinds of hypotheses have been suggested to explain the evolution of large brain size within the primates. One group of explanations emphasises the ecological function of cognitive skills, especially in large ecologically flexible species like primates (Clutton-Brock & Harvey, 1980; Gibson, 1986; Milton, 1988). The second emphasises the uniquely complex nature of primate social life, arguing for a mainly social function to intellect (Jolly, 1969; Humphrey, 1976; Kummer, 1982; Byrne & Whiten, 1988). The third type of explanation argues that neonatal brain size is constrained by maternal metabolic rates; species therefore have large brains only when maternal nutrition is on a high enough plane to allow the mother to divert spare energy into the foetus (e.g., Martin, 1981, 1984; see also Hofman, 1983*a*,*b*; Armstrong, 1985).

The third type of explanation need not concern us here for two quite different reasons. In the first place, this kind of explanation offers a purely developmental account; it essentially states that there is a limit (imposed by maternal nutrition) beyond which foetal brain size cannot grow. But it offers no explanation of any kind as to why the brain should always grow to this limit. Given that the brain is the most expensive organ of the body to maintain (it consumes approximately 20% of the body's total energy output in humans, while accounting for only 2% of adult body weight), it is evolutionarily implausible to suggest that organisms will develop large brains merely because they can do so. Natural selection rarely leads to the evolution of characters that are wholly functionless simply because they are possible. Hence, even if it were true that energetic considerations constrain brain size, a proper functional

- Discuss results at a high level
  - Goal is not an in-depth discussion
- Results will frame our discussion of social media

- The Strength of Weak Ties
  - by Mark S. Granovetter
    - [American Journal of Sociology, vol. 78 issue 6. May 1973]

The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

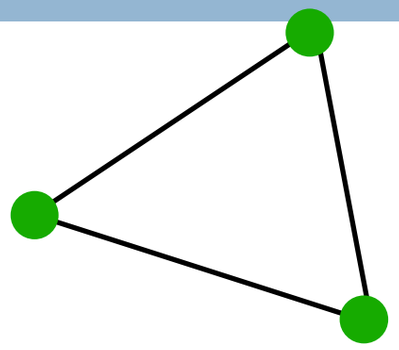
A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases.

I will argue, in this paper, that the analysis of processes in interpersonal networks provides the most fruitful micro-macro bridge. In one way or another, it is through these networks that small-scale interaction becomes translated into large-scale patterns, and that these, in turn, feed back into small groups.

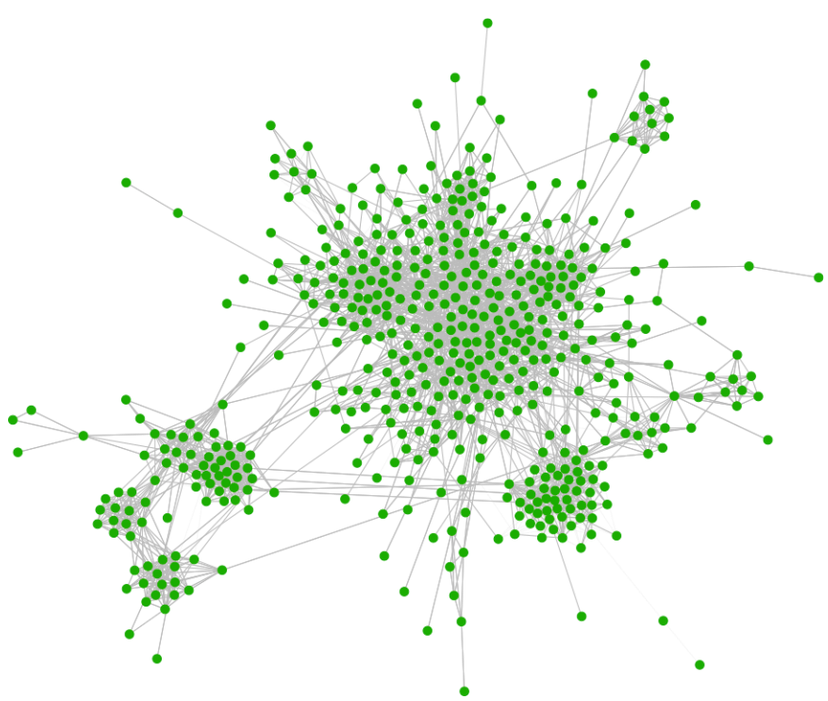
Sociometry, the precursor of network analysis, has always been curiously peripheral—invisible, really—in sociological theory. This is partly because it has usually been studied and applied only as a branch of social psychology; it is also because of the inherent complexities of precise network analysis. We have had neither the theory nor the measurement and sampling techniques to move sociometry from the usual small-group level to that of larger structures. While a number of stimulating and suggestive

[1] This paper originated in discussions with Harrison White, to whom I am indebted for many suggestions and ideas. Earlier drafts were read by Ivan Chase, James Davis, William Michelson, Nancy Lee, Peter Rossi, Charles Tilly, and an anonymous referee; their criticisms resulted in significant improvements.

“Classical” sociology



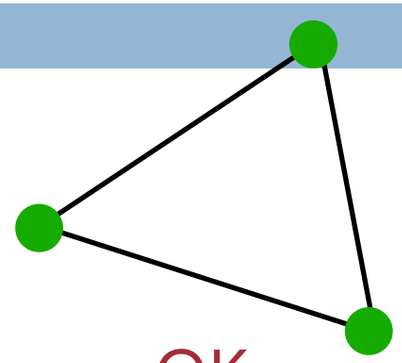
- Focused on two topics
  - Micro-level interactions within a small group
  - Macro-level patterns within a society
- “Strong” ties considered the important ones
  - Close friends, family
  - “Weak” ties considered less important



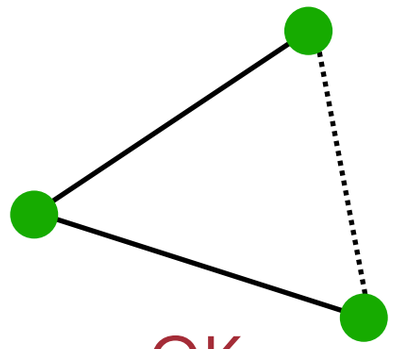
- But, mapping not understood
  - How do large-scale patterns emerge?
  - ...certain analogies to physics...

Granovetter's idea

- Construct simple model:
  - If two people have a common strong tie, they must have a tie between each other



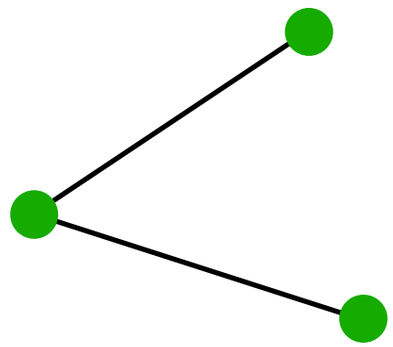
OK



OK

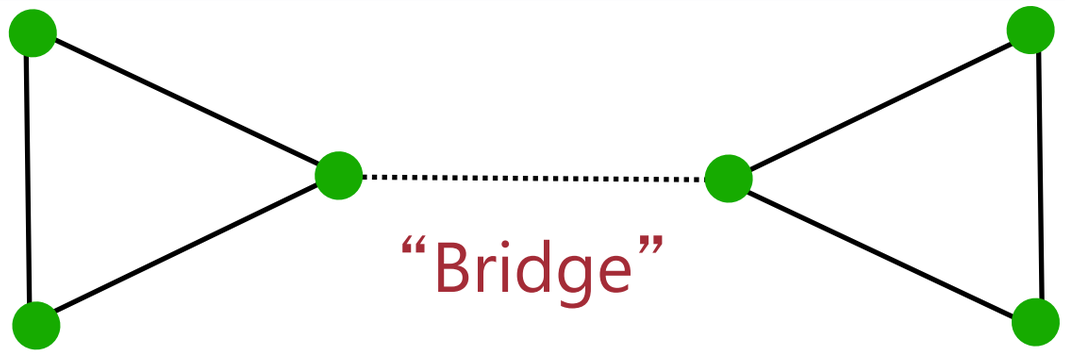
- Matches intuition from real world
  - If you have two close friends, they (at least) know each other

- What are the implications of this model?



Forbidden

Bridges



- Social networks can be divided into communities
  - Clubs, schools, employers, …
- Define a bridge as a link that is the only path between two users
- Claim: With Granovetter's assumption, bridges must be weak
  - Why?

Importance of bridges

- Bridges connect communities
  - Build up society from a set of communities

- Thus, weak ties (bridges) can help the micro → macro mapping

- Bridges must necessarily carry any new information
  - Example: People often find new jobs via weak ties
  - Societies with weak ties better able to adapt
  - Hence, the strength of weak ties

- But, what is the structure of weak ties at scale?
  - Are they really necessary for conveying information

- An Experimental Study of the Small World Problem
  - by Jeffery Travers and Stanley M
    - [Sociometry, vol.32 no. 4. 1969]

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.

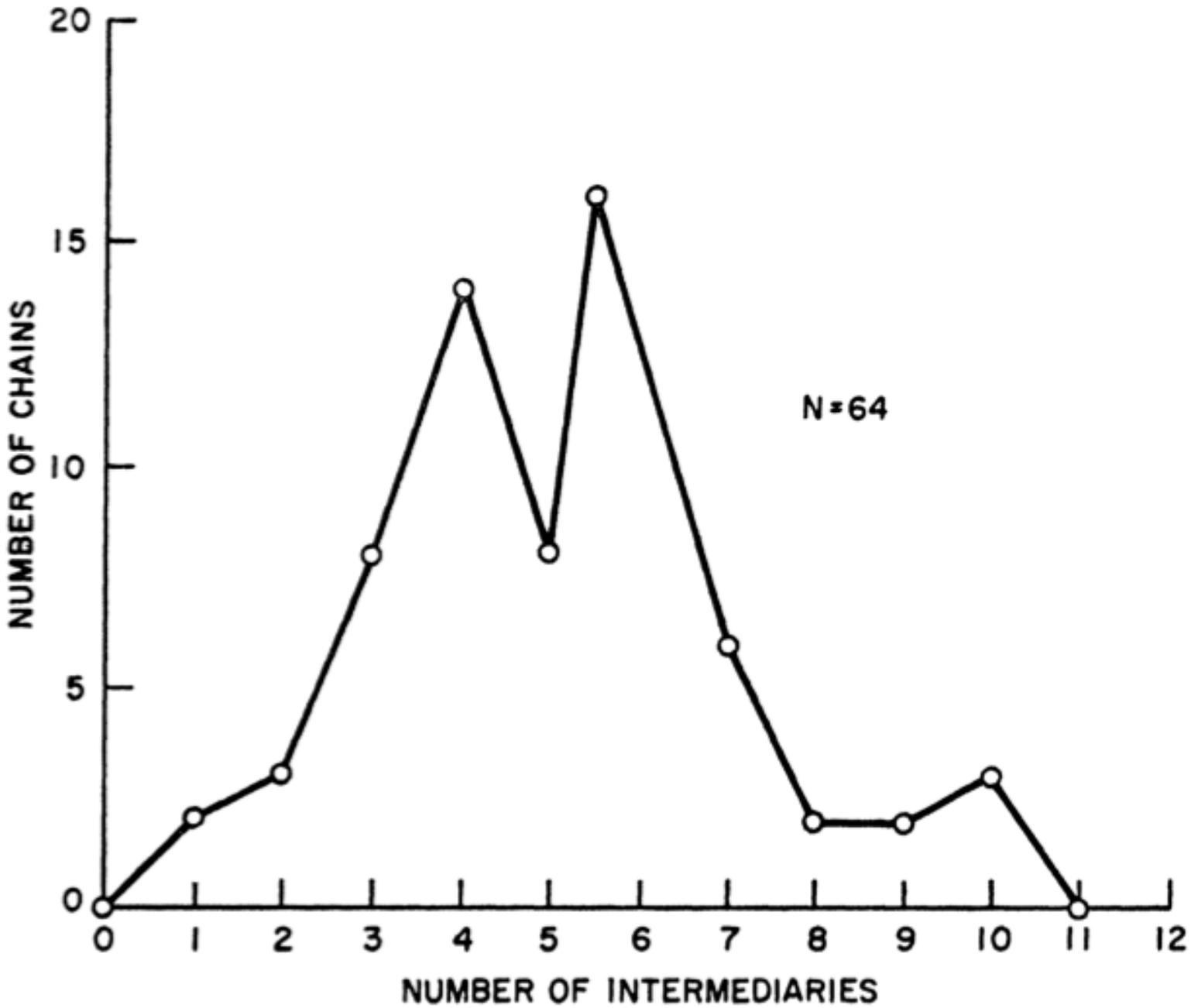
Procedure

- Selected 296 people in Nebraska and Boston
  - Mailed a packet containing instructions

- Packet specified a destination person
  - Name, address, profession, and city

- Asked to forward to someone known personally
  - Send a card back to Milgram
  - And add name to a roster
    - Why?

How long are the (successful) paths?



Implications

- Not only do short chains exist…
  - But people can find them!
  - With only local information

- Thus, social networks are navigable
  - 48% of 64 chains coalesced into 3 people
  - Important structural properties

- However, how did users "route"?
  - Did they rely on certain network properties?
  - Do shorter paths exist?

- Neocortex Size as a Constraint Group Size in Primates
  - by Robin I. M. Dunbar
    - [Journal of Human Evolution, vol. 2

R. I. M. Dunbar

Department of Anthropology, University College London, Gower St, London WC1E 6BT, U.K.

Received 3 March 1989
Revision received 18 October 1991 and accepted 2 December 1991

Keywords: behavioural ecology, grooming, brain size, body size, social intellect.

Neocortex size as a constraint on group size in primates

Two general kinds of theory (one ecological and one social) have been advanced to explain the fact that primates have larger brains and greater cognitive abilities than other animals. Data on neocortex volume, group size and a number of behavioural ecology variables are used to test between the various theories. Group size is found to be a function of relative neocortical volume, but the ecological variables are not. This is interpreted as evidence in favour of the social intellect theory and against the ecological theories. It is suggested that the number of neocortical neurons limits the organism's information-processing capacity and that this then limits the number of relationships that an individual can monitor simultaneously. When a group's size exceeds this limit, it becomes unstable and begins to fragment. This then places an upper limit on the size of groups which any given species can maintain as cohesive social units through time. The data suggest that the information overload occurs in terms of the structure of relationships within tightly bonded grooming cliques rather than in terms of the total number of dyads within the group as a whole that an individual has to monitor. It thus appears that, among primates, large groups are created by welding together sets of smaller grooming cliques. One implication of these results is that, since the actual group size will be determined by the ecological characteristics of the habitat in any given case, species will only be able to invade habitats that require larger groups than their current limit if they evolve larger neocortices.

Journal of Human Evolution (1992) **20,** 469–493

Introduction

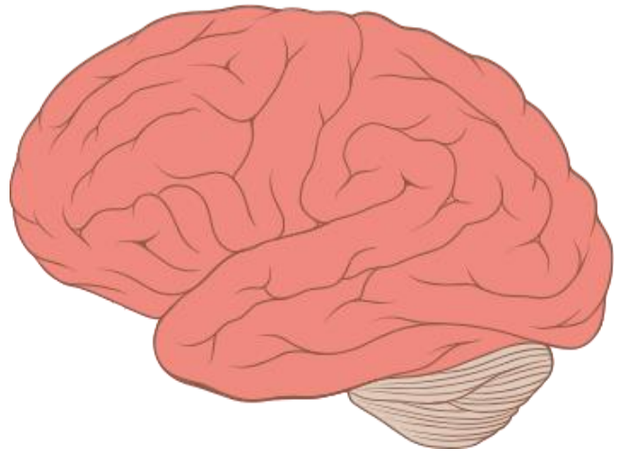
Primates, as a group, are characterised by having unusually large brains for their body size (Jerison 1973). Implicitly or explicitly, it has usually been assumed that large relative brain size correlates with these animals' greater cognitive ability. Three general kinds of hypotheses have been suggested to explain the evolution of large brain size within the primates. One group of explanations emphasises the ecological function of cognitive skills, especially in large ecologically flexible species like primates (Clutton-Brock & Harvey, 1980; Gibson, 1986; Milton, 1988). The second emphasises the uniquely complex nature of primate social life, arguing for a mainly social function to intellect (Jolly, 1969; Humphrey, 1976; Kummer, 1982; Byrne & Whiten, 1988). The third type of explanation argues that neonatal brain size is constrained by maternal metabolic rates; species therefore have large brains only when maternal nutrition is on a high enough plane to allow the mother to divert spare energy into the foetus (e.g., Martin, 1981, 1984; see also Hofman, 1983*a*,*b*; Armstrong, 1985).

The third type of explanation need not concern us here for two quite different reasons. In the first place, this kind of explanation offers a purely developmental account; it essentially states that there is a limit (imposed by maternal nutrition) beyond which foetal brain size cannot grow. But it offers no explanation of any kind as to why the brain should always grow to this limit. Given that the brain is the most expensive organ of the body to maintain (it consumes approximately 20% of the body's total energy output in humans, while accounting for only 2% of adult body weight), it is evolutionarily implausible to suggest that organisms will develop large brains merely because they can do so. Natural selection rarely leads to the evolution of characters that are wholly functionless simply because they are possible. Hence, even if it were true that energetic considerations constrain brain size, a proper functional

0047–2484/92/060469+25 $03.00/0

Neocortex

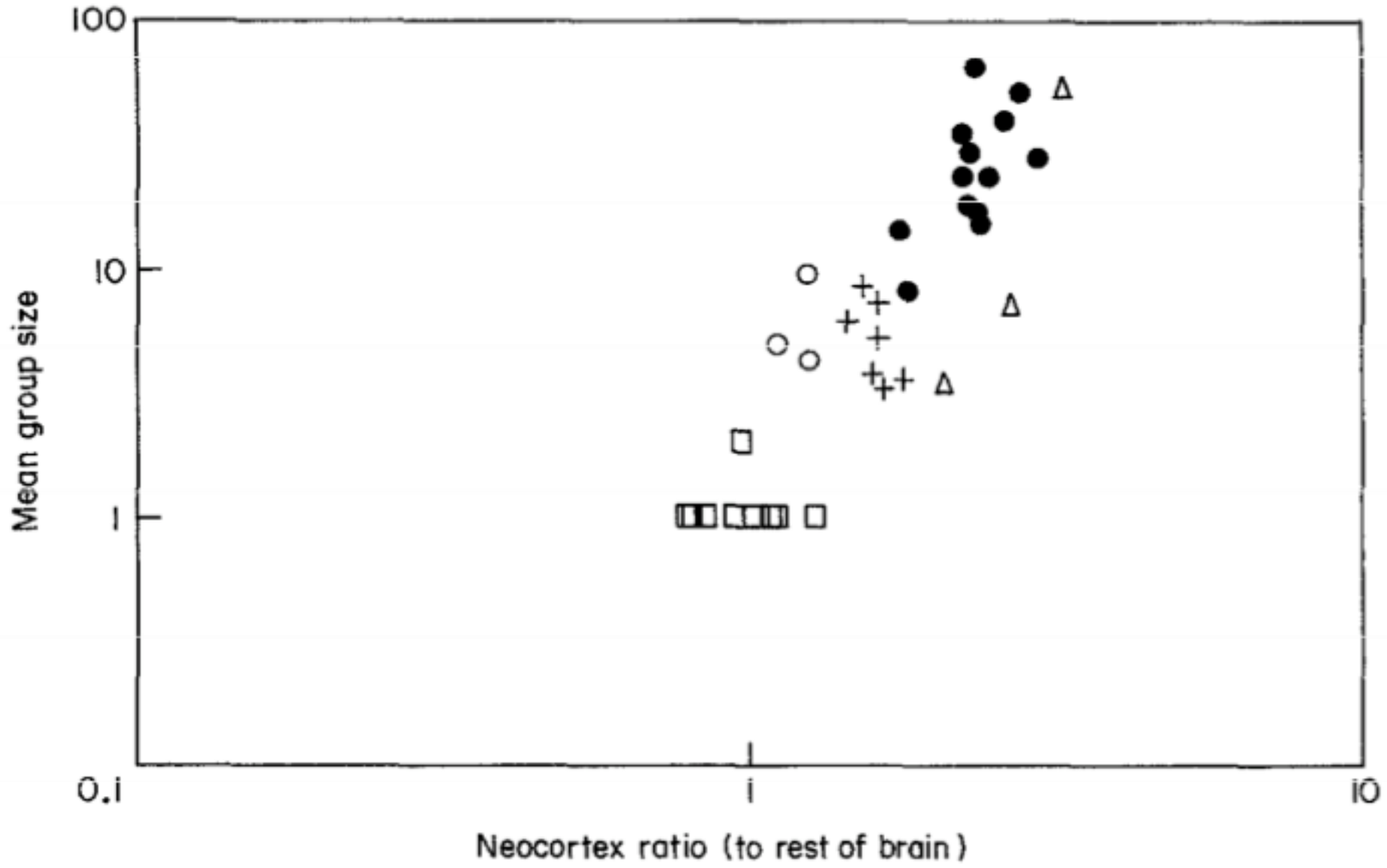
- Part of the brain of mammals, involved in
  - Sensory perception
  - Motor commands
  - Spatial reasoning
  - Thought and language
  - Social interactions



Neocortex

- Theory: large brain size due to "social" nature of primates
  - Measure "social" level by looking at typical group size
  - If true, then brain size should correlate with being "social"

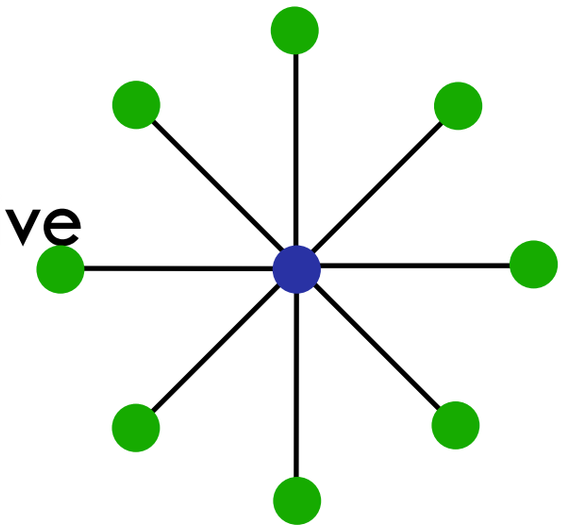
Neocortex size and group size



- Strong correlation observed
  - Holds across many species of primates

Implications

- Each individual can only maintain so many relationships
  - Bounded by neocortex size



- Not just the number of relationship, it is the intensive with which a small number of key "friendships"
  - Who likes who, who doesn't, etc

- Is this true for humans?
  - Social groups are less well-defined
  - Dunbar predicts value of 150 from neocortex size

- What about different relationship types?
  - What is the variance across individuals?

Social science primer: Summary

- Doing this sort of work takes significant effort!

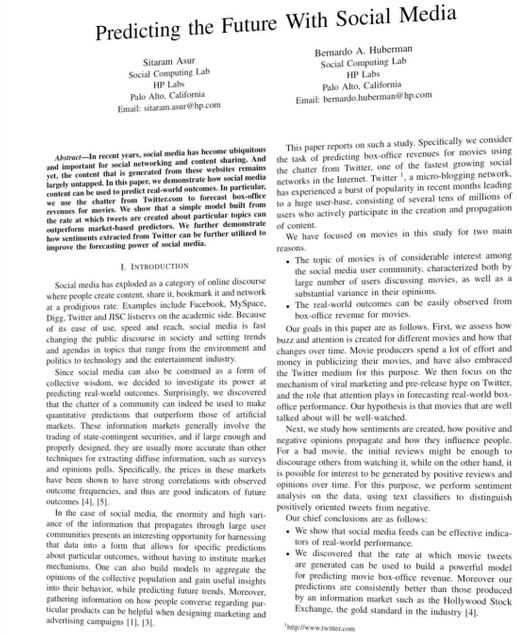
- Key results:
  - Network structure influenced by strong/weak links
  - Networks have (navigable) short paths
  - Expected bound on degree for each node

- Do results hold for at large scale?
  - Or, for social media at all?

- What social science questions can we answer with social media?

LEVERAGING SOCIAL MEDIA

Two papers on leveraging social media



Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

Abstract—**In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.**

I. Introduction

Social media has exploded as a category of online discourse where people create content, share it, bookmark it and network at a prodigious rate. Examples include Facebook, MySpace, Digg, Twitter and JISC listservs on the academic side. Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry.

Since social media can also be construed as a form of collective wisdom, we decided to investigate its power at predicting real-world outcomes. Surprisingly, we discovered that the chatter of a community can indeed be used to make quantitative predictions that outperform those of artificial markets. These information markets generally involve the trading of state-contingent securities, and if large enough and properly designed, they are usually more accurate than other techniques for extracting diffuse information, such as surveys and opinions polls. Specifically, the prices in these markets have been shown to have strong correlations with observed outcome frequencies, and thus are good indicators of future outcomes [4], [5].

In the case of social media, the enormity and high variance of the information that propagates through large user communities presents an interesting opportunity for harnessing that data into a form that allows for specific predictions about particular outcomes, without having to institute market mechanisms. One can also build models to aggregate the opinions of the collective population and gain useful insights into their behavior, while predicting future trends. Moreover, gathering information on how people converse regarding particular products can be helpful when designing marketing and advertising campaigns [1], [3].

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of content.

We have focused on movies in this study for two main reasons.

- The topic of movies is of considerable interest among the social media user community, characterized both by large number of users discussing movies, as well as a substantial variance in their opinions.
- The real-world outcomes can be easily observed from box-office revenue for movies.

Our goals in this paper are as follows. First, we assess how buzz and attention is created for different movies and how that changes over time. Movie producers spend a lot of effort and money in publicizing their movies, and have also embraced the Twitter medium for this purpose. We then focus on the mechanism of viral marketing and pre-release hype on Twitter, and the role that attention plays in forecasting real-world box-office performance. Our hypothesis is that movies that are well talked about will be well-watched.

Next, we study how sentiments are created, how positive and negative opinions propagate and how they influence people. For a bad movie, the initial reviews might be enough to discourage others from watching it, while on the other hand, it is possible for interest to be generated by positive reviews and opinions over time. For this purpose, we perform sentiment analysis on the data, using text classifiers to distinguish positively oriented tweets from negative.

Our chief conclusions are as follows:

- We show that social media feeds can be effective indicators of real-world performance.
- We discovered that the rate at which movie tweets are generated can be used to build a powerful model for predicting movie box-office revenue. Moreover our predictions are consistently better than those produced by an information market such as the Hollywood Stock Exchange, the gold standard in the industry [4].

[1]http://www.twitter.com

Meme-tracking and the Dynamics of the News Cycle

Jure Leskovec*† Lars Backstrom* Jon Kleinberg*
*Cornell University †Stanford University
jure@cs.stanford.edu lars@cs.cornell.edu kleinber@cs.cornell.edu

ABSTRACT

Tracking new topics, ideas, and "memes" across the Web has been an issue of considerable interest. Recent work has developed methods for tracking topic shifts over long time scales, as well as abrupt spikes in the appearance of particular named entities. However, these approaches are less well suited to the identification of content that spreads widely and then fades over time scales on the order of days — the time scale at which we perceive news and events.

We develop a framework for tracking short, distinctive phrases that travel relatively intact through on-line text; developing scalable algorithms for clustering textual variants of such phrases, we identify a broad class of memes that exhibit wide spread and rich variation on a daily basis. As our principal domain of study, we show how such a meme-tracking approach can provide a coherent representation of the *news cycle* — the daily rhythms in the news media that have long been the subject of qualitative interpretation but have never been captured accurately enough to permit actual quantitative analysis. We tracked 1.6 million mainstream media sites and blogs over a period of three months with the total of 90 million articles and we find a set of novel and persistent temporal patterns in the news cycle. In particular, we observe a typical lag of 2.5 hours between the peaks of attention to a phrase in the news media and in blogs respectively, with divergent behavior around the overall peak and a "heartbeat"-like pattern in the handoff between news and blogs. We also develop and analyze a mathematical model for the kinds of temporal variation that the system exhibits.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database applications—*Data mining*

General Terms: Algorithms; Experimentation.

Keywords: Meme-tracking, Blogs, News media, News cycle, Information cascades, Information diffusion, Social networks

1. INTRODUCTION

A growing line of research has focused on the issues raised by the diffusion and evolution of highly dynamic on-line information, particularly the problem of tracking topics, ideas, and "memes" as they evolve over time and spread across the web. Prior work has identified two main approaches to this problem, which have been successful at two correspondingly different extremes of it. Probabilistic term mixtures have been successful at identifying long-range trends in general topics over time [5, 7, 16, 17, 30, 31]. At the other extreme, identifying hyperlinks between blogs and extracting rare named entities has been used to track short information cascades through the blogosphere [3, 14, 20, 23]. However, between these two extremes lies much of the temporal and textual range over which propagation on the web and between people typically occurs, through the continuous interaction of news, blogs, and websites on a daily basis. Intuitively, short units of text, short phrases, and "memes" that act as signatures of topics and events propagate and diffuse over the web, from mainstream media to blogs, and vice versa. This is exactly the focus of our study here.

Moreover, it is at this intermediate temporal and textual granularity of memes and phrases that people experience news and current events. A succession of story lines that evolve and compete for attention within a relatively stable set of broader topics collectively produces an effect that commentators refer to as the *news cycle*. Tracking dynamic information at this temporal and topical resolution has proved difficult, since the continuous appearance, growth, and decay of new story lines takes place without significant shifts in the overall vocabulary; in general, this process can also not be closely aligned with the appearance and disappearance of specific named entities (or hyperlinks) in the text. As a result, while the dynamics of the news cycle has been a subject of intense interest to researchers in media and the political process, the focus has been mainly qualitative, with a corresponding lack of techniques for undertaking quantitative analysis of the news cycle as a whole.

Our approach to meme-tracking, with applications to the news cycle. Here we develop a method for tracking units of information as they spread over the web. Our approach is the first to scalably identify short distinctive phrases that travel relatively intact through on-line text as it evolves over time. Thus, for the first time at a large scale, we are able to automatically identify and actually "see" such textual elements and study them in a massive dataset providing essentially complete coverage of on-line mainstream and blog media. Working with phrases naturally interpolates between the two extremes of topic models on the one hand and named entities on the other. First, the set of distinctive phrases shows significant diversity over short periods of time, even as the broader vocabulary remains relatively stable. As a result, they can be used to dissect a general topic into a large collection of threads or memes that vary from day to day. Second, such distinctive phrases are abundant, and therefore are rich enough to act as "tracers" for a large collection of memes; we therefore do not have to restrict attention to the much smaller collection of memes that happen to be associated with the appearance and disappearance of a single named entity.

From an algorithmic point of view, we consider these distinctive phrases to act as the analogue of "genetic signatures" for different

- Cover two topics
  - Tracking information flow
  - Applying social media to real-world problems

- Meme-tracking and the Dynam of the News Cycle
  - by Jure Leskovec, Lars Backstror Jon Kleinberg
- [Proceedings of KDD 2009]

Meme-tracking and the Dynamics of the News Cycle

Jure Leskovec*† Lars Backstrom* Jon Kleinberg*
*Cornell University †Stanford University
jure@cs.stanford.edu lars@cs.cornell.edu kleinber@cs.cornell.edu

ABSTRACT

Tracking new topics, ideas, and "memes" across the Web has been an issue of considerable interest. Recent work has developed methods for tracking topic shifts over long time scales, as well as abrupt spikes in the appearance of particular named entities. However, these approaches are less well suited to the identification of content that spreads widely and then fades over time scales on the order of days — the time scale at which we perceive news and events.

We develop a framework for tracking short, distinctive phrases that travel relatively intact through on-line text; developing scalable algorithms for clustering textual variants of such phrases, we identify a broad class of memes that exhibit wide spread and rich variation on a daily basis. As our principal domain of study, we show how such a meme-tracking approach can provide a coherent representation of the *news cycle* — the daily rhythms in the news media that have long been the subject of qualitative interpretation but have never been captured accurately enough to permit actual quantitative analysis. We tracked 1.6 million mainstream media sites and blogs over a period of three months with the total of 90 million articles and we find a set of novel and persistent temporal patterns in the news cycle. In particular, we observe a typical lag of 2.5 hours between the peaks of attention to a phrase in the news media and in blogs respectively, with divergent behavior around the overall peak and a "heartbeat"-like pattern in the handoff between news and blogs. We also develop and analyze a mathematical model for the kinds of temporal variation that the system exhibits.

Categories and Subject Descriptors: H.2.8 **[Database Management]**: Database applications—*Data mining*

General Terms: Algorithms; Experimentation.

Keywords: Meme-tracking, Blogs, News media, News cycle, Information cascades, Information diffusion, Social networks

1. INTRODUCTION

A growing line of research has focused on the issues raised by the diffusion and evolution of highly dynamic on-line information, particularly the problem of tracking topics, ideas, and "memes" as they evolve over time and spread across the web. Prior work has identified two main approaches to this problem, which have been successful at two correspondingly different extremes of it. Probabilistic term mixtures have been successful at identifying long-range trends in general topics over time [5, 7, 16, 17, 30, 31]. At the other extreme, identifying hyperlinks between blogs and extracting rare named entities has been used to track short information cascades through the blogosphere [3, 14, 20, 23]. However, between these two extremes lies much of the temporal and textual range over which propagation on the web and between people typically occurs, through the continuous interaction of news, blogs, and websites on a daily basis. Intuitively, short units of text, short phrases, and "memes" that act as signatures of topics and events propagate and diffuse over the web, from mainstream media to blogs, and vice versa. This is exactly the focus of our study here.

Moreover, it is at this intermediate temporal and textual granularity of memes and phrases that people experience news and current events. A succession of story lines that evolve and compete for attention within a relatively stable set of broader topics collectively produces an effect that commentators refer to as the *news cycle*. Tracking dynamic information at this temporal and topical resolution has proved difficult, since the continuous appearance, growth, and decay of new story lines takes place without significant shifts in the overall vocabulary; in general, this process can also not be closely aligned with the appearance and disappearance of specific named entities (or hyperlinks) in the text. As a result, while the dynamics of the news cycle has been a subject of intense interest to researchers in media and the political process, the focus has been mainly qualitative, with a corresponding lack of techniques for undertaking quantitative analysis of the news cycle as a whole.

Our approach to meme-tracking, with applications to the news cycle. Here we develop a method for tracking units of information as they spread over the web. Our approach is the first to scalably identify short distinctive phrases that travel relatively intact through on-line text as it evolves over time. Thus, for the first time at a large scale, we are able to automatically identify and actually "see" such textual elements and study them in a massive dataset providing essentially complete coverage of on-line mainstream and blog media. Working with phrases naturally interpolates between the two extremes of topic models on the one hand and named entities on the other. First, the set of distinctive phrases shows significant diversity over short periods of time, even as the broader vocabulary remains relatively stable. As a result, they can be used to dissect a general topic into a large collection of threads or memes that vary from day to day. Second, such distinctive phrases are abundant, and therefore are rich enough to act as "tracers" for a large collection of memes; we therefore do not have to restrict attention to the much smaller collection of memes that happen to be associated with the appearance and disappearance of a single named entity.

From an algorithmic point of view, we consider these distinctive phrases to act as the analogue of "genetic signatures" for different

Leveraging social media

- Networks are used to spread information
  - Can social media shed light on information flow through society?

- Focus on news media
  - How do people find out about news?

- Who "finds" stories?
  - What role does the media/social web play?
  - How do they influence each other?

- This paper: Can social media shed light on information flow?

Memes

- Meme: Unit of culture
  - Coined by Dawkins
  - Describes evolution of culture

- Internet examples: Rickroll, LOLCat, FAIL

- Focus on memes
  - Entities (Obama) too course-grained
  - Common sequences (web 2.0) too noisy
  - Hyperlinks too fine-grained

- Use quotes to extract memes
  - "...palling around with terrorists..."

Data collected

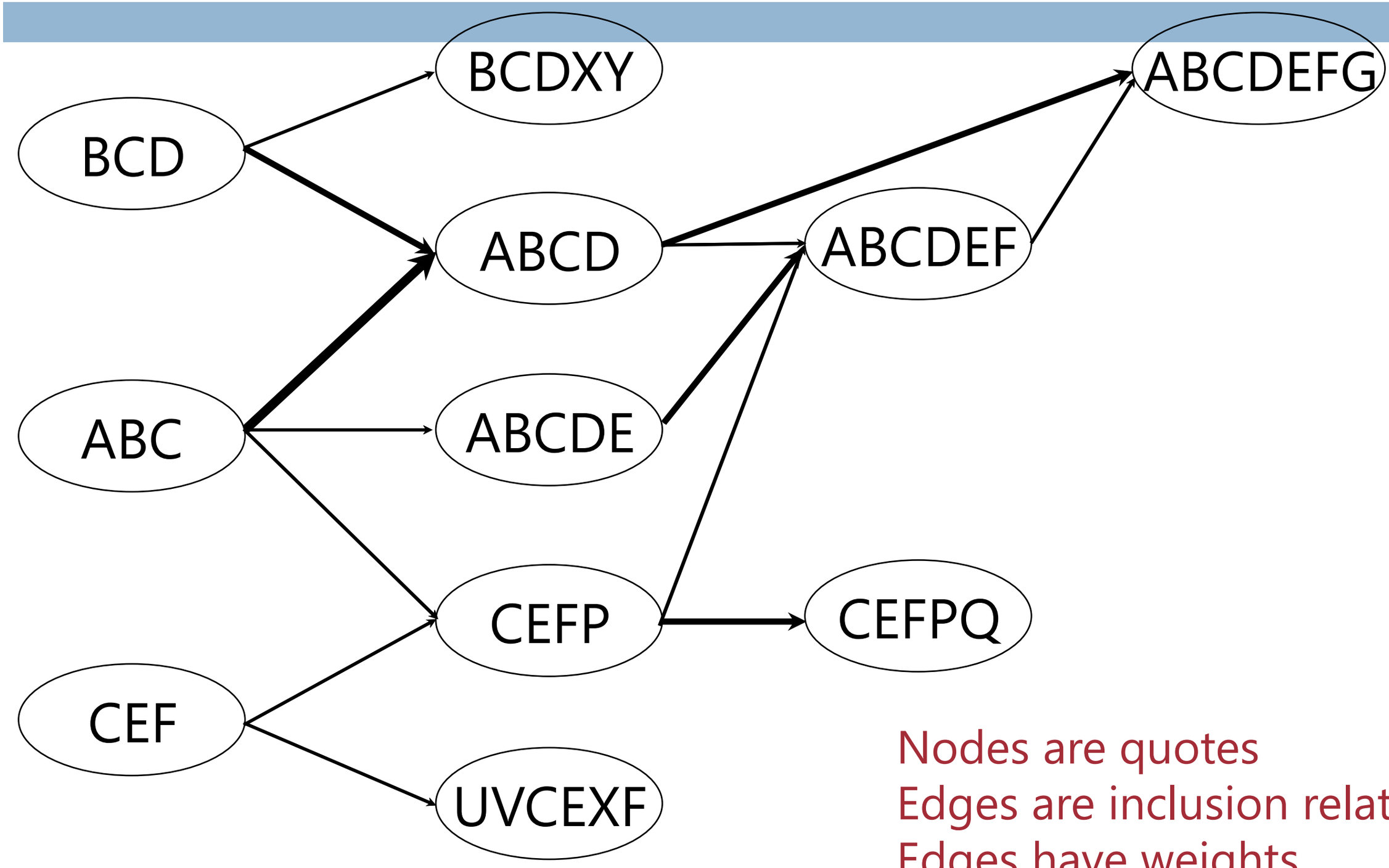
- Use dataset from spinn3r.com
  - August - October 2008
  - 90 million documents (blog entries/news stories)
    - 1.65 million sites
  - 112 million quotes

spinn3r

- Challenge: Quotes mutate
  - “...terrorists who would target their own country...”
  - “...terrorists who targeted their own country...”
  - “...terrorists who target their own country...”
  - “...terrorists who would bomb their own country...”

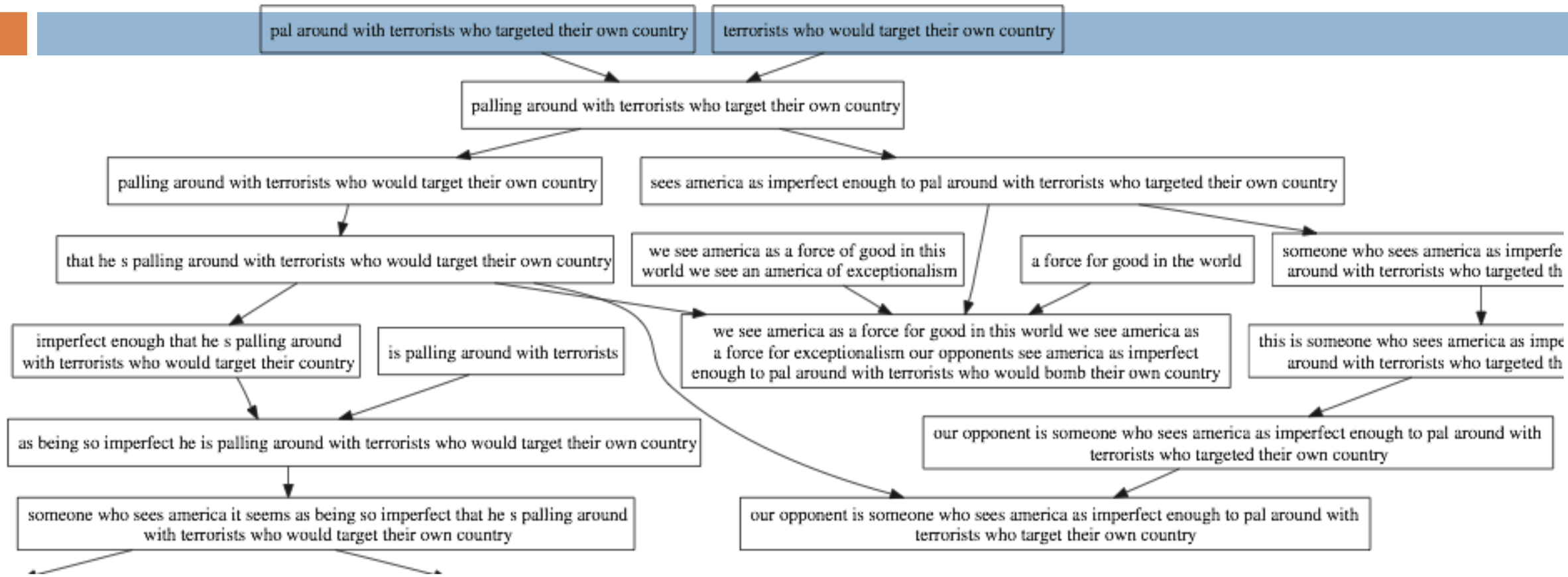
- How to determine which quotes are the same?

Clustering quotes



Nodes are quotes
Edges are inclusion relations
Edges have weights

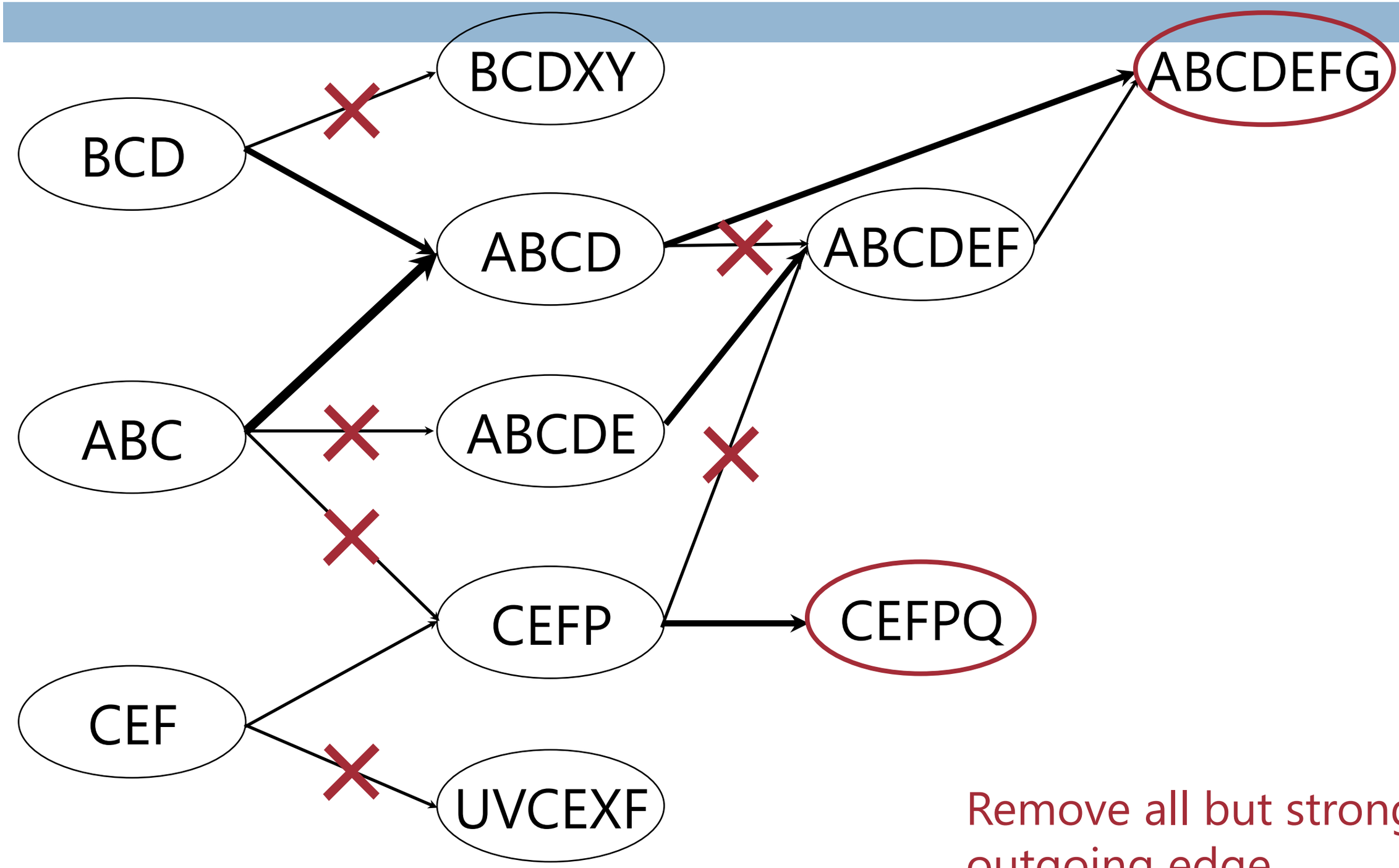
Example of cluster



- All based on Sarah Palin's terrorists quote:
  - "Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that he's palling around with terrorists who would target their own country."

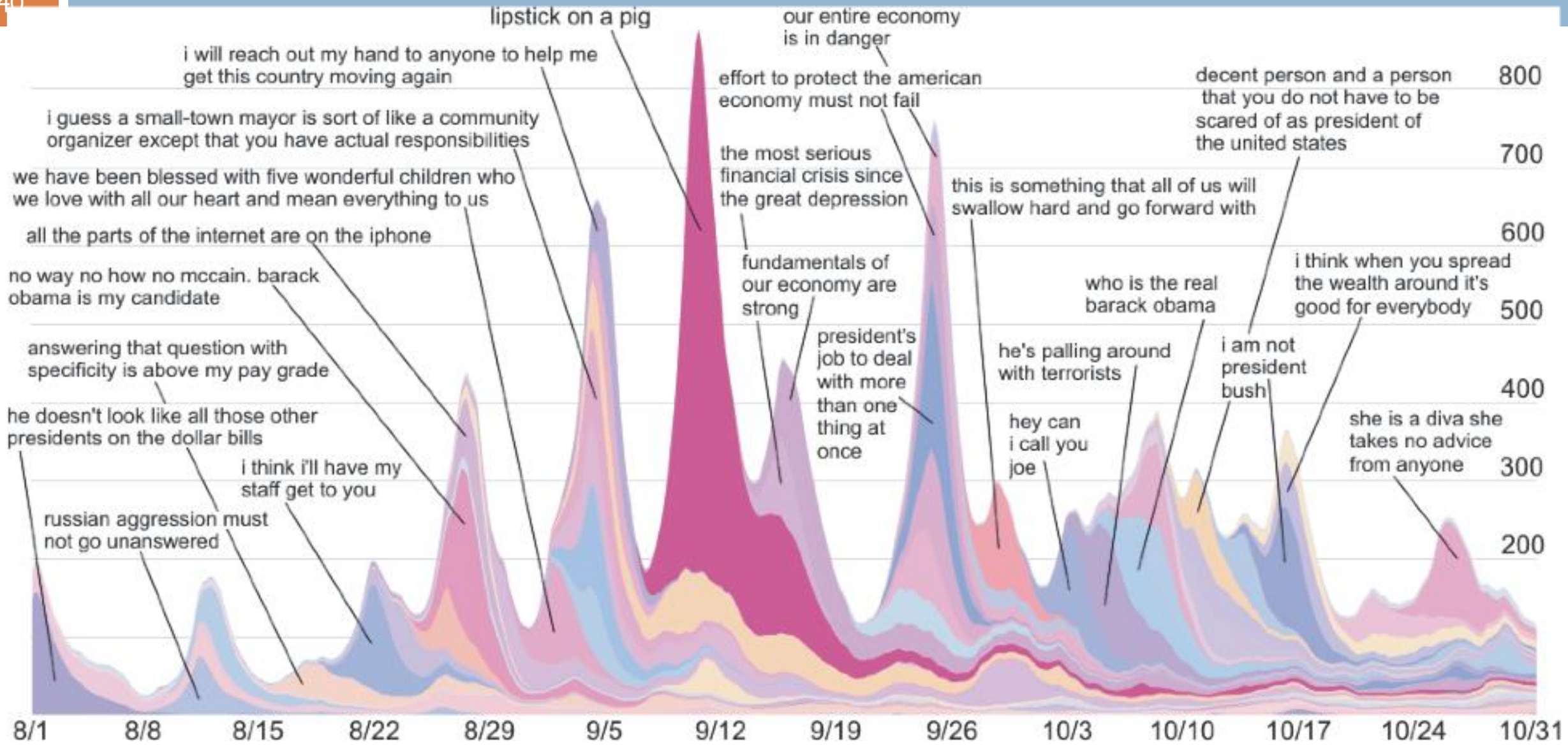
- How to reduce to a single meme?

Solution: Create a DAG



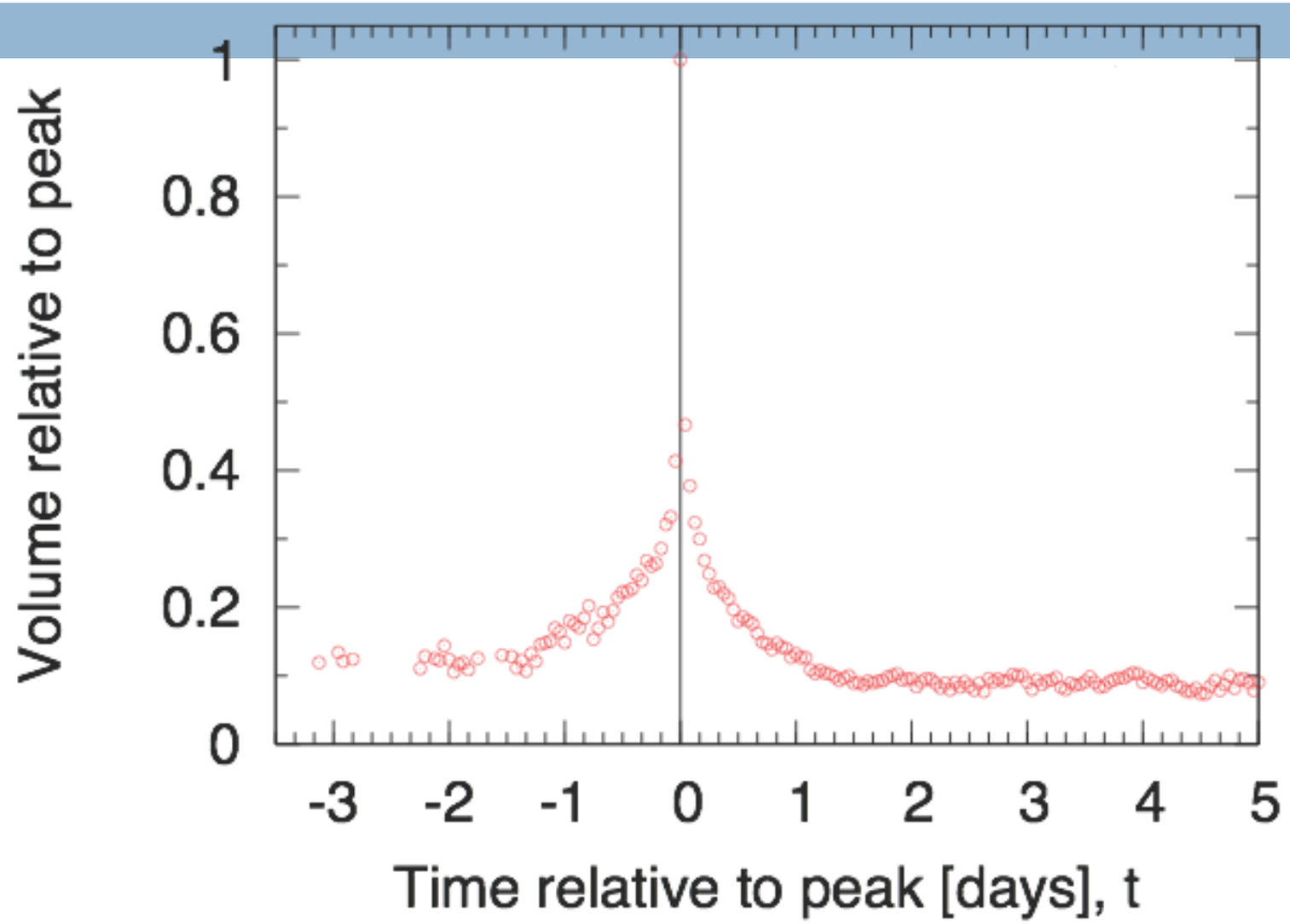
Remove all but strongest outgoing edge

Resulting memes



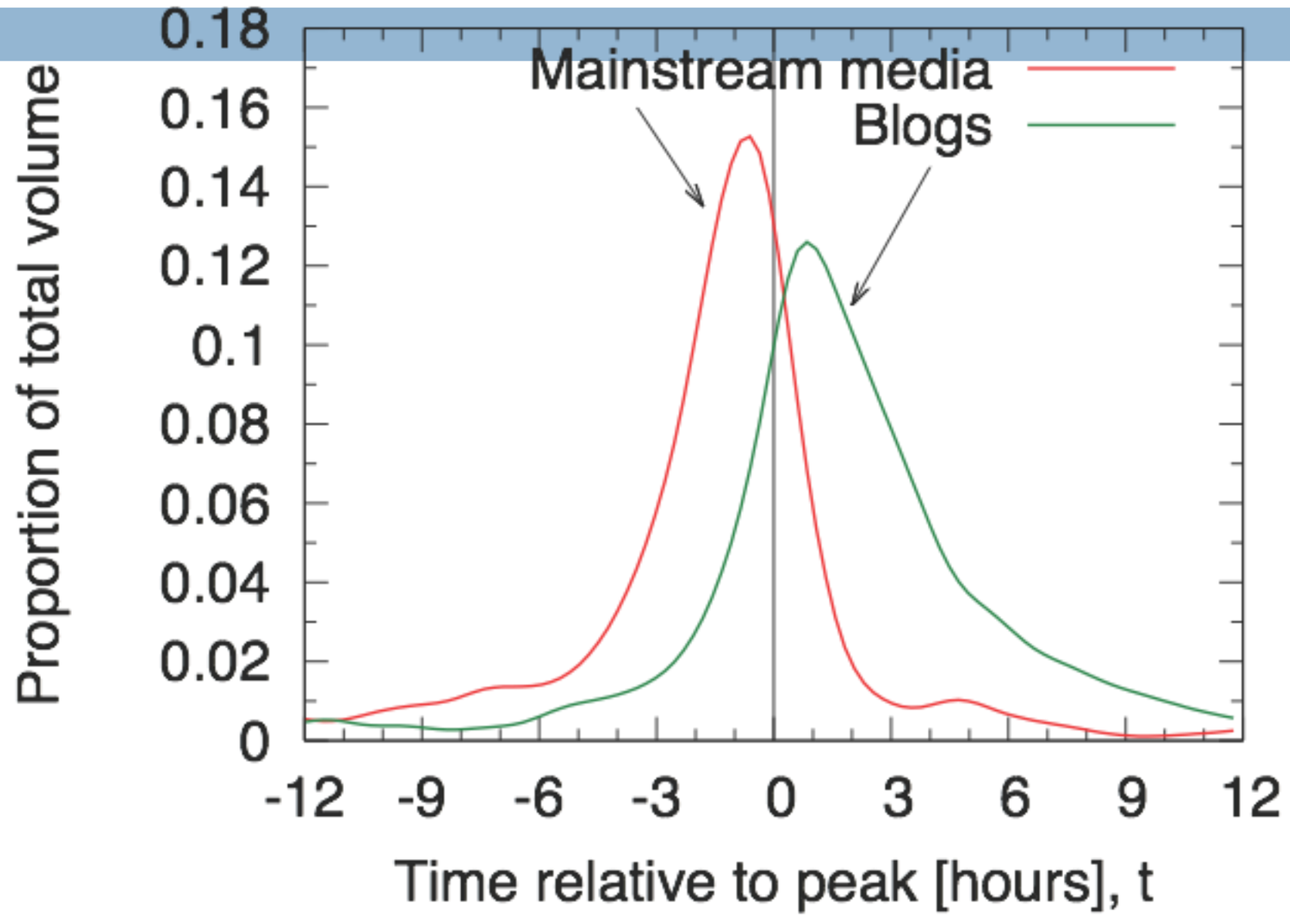
- Spikes show nature of 24-hour news cycle
  - Memes quickly enter and leave collective conscience

Tracking memes



- First, determine "peak" intensity of each meme
  - Distinct peak present

Where do the memes come from?



- Second, track where articles come from
  - Media peak is 2.5 hours before blog peak
  - Blog volume persists much longer

Summary

- Can social media shed light on information flow?

- Collected data on over 90 million documents
  - Unprecedented scale
- Found interesting interaction between media and blogs
  - Media has short attention span
    - But causes peak intensity
  - Blogs have more persistent volume

- Predicting the Future With Social Media
  - by Sitaram Asur and Bernardo A. Huberman[Arxiv 1003.2699]

Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

***Abstract*—In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.**

I. Introduction

Social media has exploded as a category of online discourse where people create content, share it, bookmark it and network at a prodigious rate. Examples include Facebook, MySpace, Digg, Twitter and JISC listservs on the academic side. Because of its ease of use, speed and reach, social media is fast changing the public discourse in society and setting trends and agendas in topics that range from the environment and politics to technology and the entertainment industry.

Since social media can also be construed as a form of collective wisdom, we decided to investigate its power at predicting real-world outcomes. Surprisingly, we discovered that the chatter of a community can indeed be used to make quantitative predictions that outperform those of artificial markets. These information markets generally involve the trading of state-contingent securities, and if large enough and properly designed, they are usually more accurate than other techniques for extracting diffuse information, such as surveys and opinions polls. Specifically, the prices in these markets have been shown to have strong correlations with observed outcome frequencies, and thus are good indicators of future outcomes [4], [5].

In the case of social media, the enormity and high variance of the information that propagates through large user communities presents an interesting opportunity for harnessing that data into a form that allows for specific predictions about particular outcomes, without having to institute market mechanisms. One can also build models to aggregate the opinions of the collective population and gain useful insights into their behavior, while predicting future trends. Moreover, gathering information on how people converse regarding particular products can be helpful when designing marketing and advertising campaigns [1], [3].

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of content.

We have focused on movies in this study for two main reasons.

- The topic of movies is of considerable interest among the social media user community, characterized both by large number of users discussing movies, as well as a substantial variance in their opinions.
- The real-world outcomes can be easily observed from box-office revenue for movies.

Our goals in this paper are as follows. First, we assess how buzz and attention is created for different movies and how that changes over time. Movie producers spend a lot of effort and money in publicizing their movies, and have also embraced the Twitter medium for this purpose. We then focus on the mechanism of viral marketing and pre-release hype on Twitter, and the role that attention plays in forecasting real-world box-office performance. Our hypothesis is that movies that are well talked about will be well-watched.

Next, we study how sentiments are created, how positive and negative opinions propagate and how they influence people. For a bad movie, the initial reviews might be enough to discourage others from watching it, while on the other hand, it is possible for interest to be generated by positive reviews and opinions over time. For this purpose, we perform sentiment analysis on the data, using text classifiers to distinguish positively oriented tweets from negative.

Our chief conclusions are as follows:

- We show that social media feeds can be effective indicators of real-world performance.
- We discovered that the rate at which movie tweets are generated can be used to build a powerful model for predicting movie box-office revenue. Moreover our predictions are consistently better than those produced by an information market such as the Hollywood Stock Exchange, the gold standard in the industry [4].

[1] http://www.twitter.com

Social media and communication

- Social media enables communication
  - Facebook wall
  - Orkut scraps
  - Twitter tweets

- Essentially, we have microphone above the world
  - Have complete conversations for huge group of users
  - Can access collective wisdom

- Can we extract information from these conversations?
  - In aggregate?

This paper: twitter + movies

- Focus on twitter
  - Most data is publicly available
  - Messages are short

- Can we use twitter to predict the future?
  - Focus on box-office returns for movies
    - Relatively short term (~3 week window/movie)
  - Existing techniques to compare against
    - Gold standard is Hollywood Stock Exchange

Hollywood stock exchange (HSX)

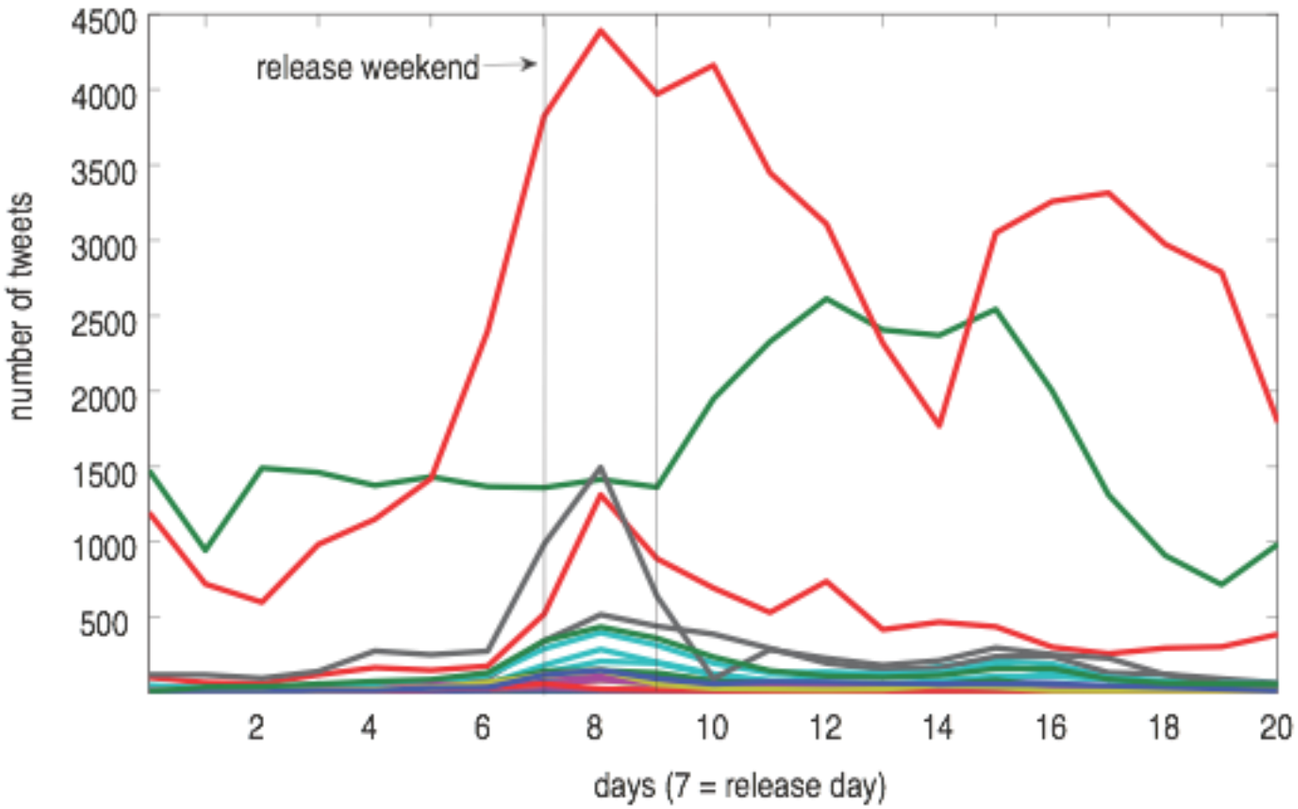
- Example of a prediction market
  - Uses play money
- Can buy movie stocks
  - Each H$ = $1M US gross take
- Each movie has a listed delist date
  - 4 weeks after open, cashed out
  - Value is US gross take
- Surprisingly accurate
  - 32 of 39 Oscar nominees in 2007
  - 7 of 8 eventual winners

Can we use social media?

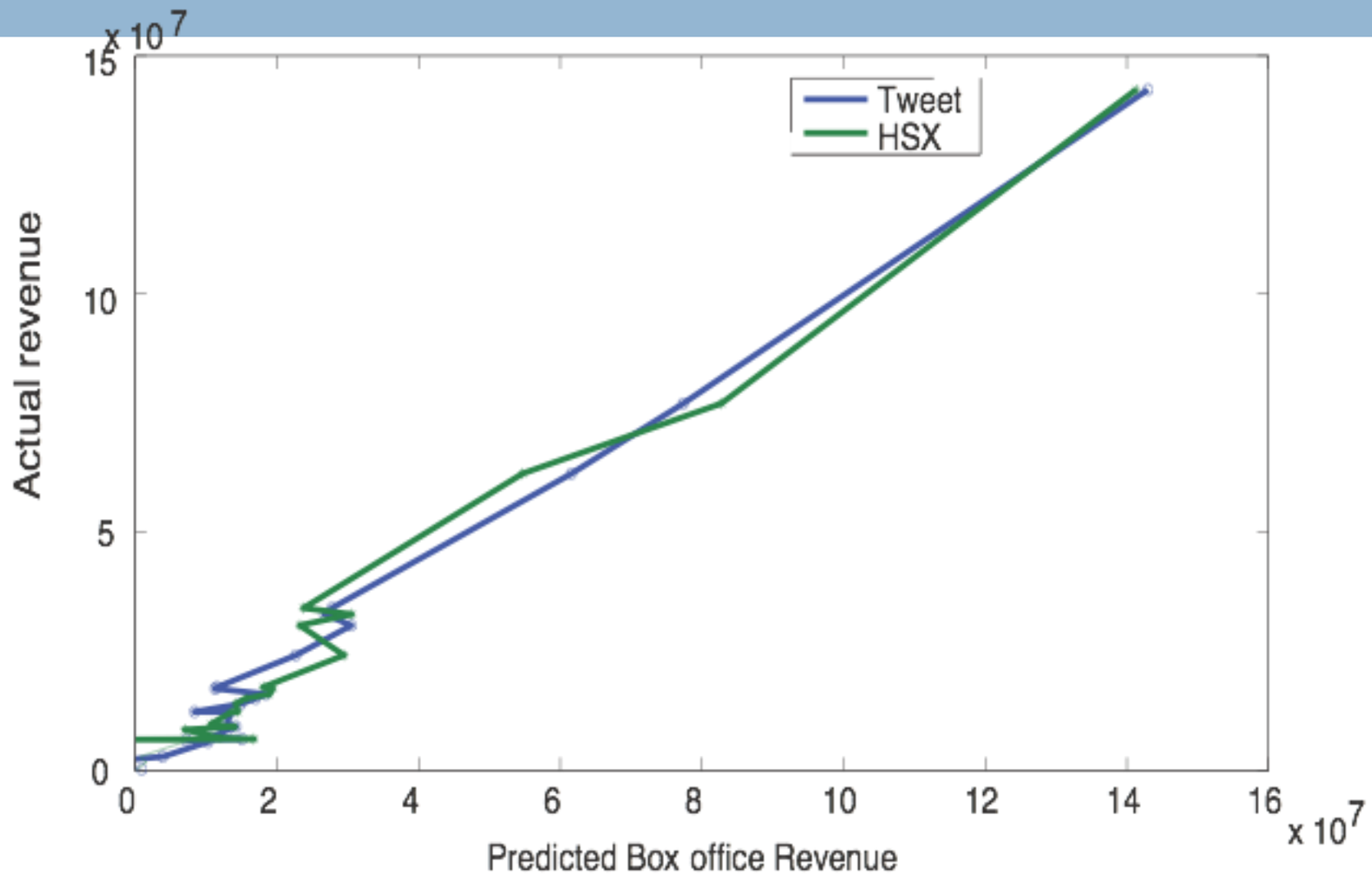
- Focus on mentions of 24 movies on twitter
  - Armored, Avatar, The Blind Side, The Book of Eli, Day breakers, Dear John, Did You Hear About The Morgans, Edge Of Darkness, Extraordinary Measures, From Paris With Love, The Imaginarium of Dr Parnassus, Invictus, Leap Year, LegionTwilight : New Moon, Pirate Radio, Princess And The Frog, Sherlock Holmes, Spy Next Door, The Crazies Tooth, Fairy Transylmania, When In Rome, Youth, In Revolt
- Obtained data by searching repeatedly
  - Three weeks around release date
  - Most activity in this period
  - Most money made in this period
- Total of 2.89M tweets

Making predictions

- Busiest time is around release
  - Promotions, advertising, …
- Opening weekend makes most money
- Predict take by looking at pre-release tweet rate
  - How many tweets before open?
  - Compare to HSX



How accurate are the predictions?



- Very accurate!
  - Coefficient of determination (R^2) is 0.973
  - Versus 0.965 for HSX

Summary

- First look at using social media for prediction
- Relatively simple approach, naïve predictor
  - Simply looking at number of mentions before release
  - Outperformed existing gold standard
- What else can we use social media to predict?
  - Stock markets?
- But unclear causality
  - Do movie studios only promote movies they expect to be a hit?