SOSC 4300/5500: Networks

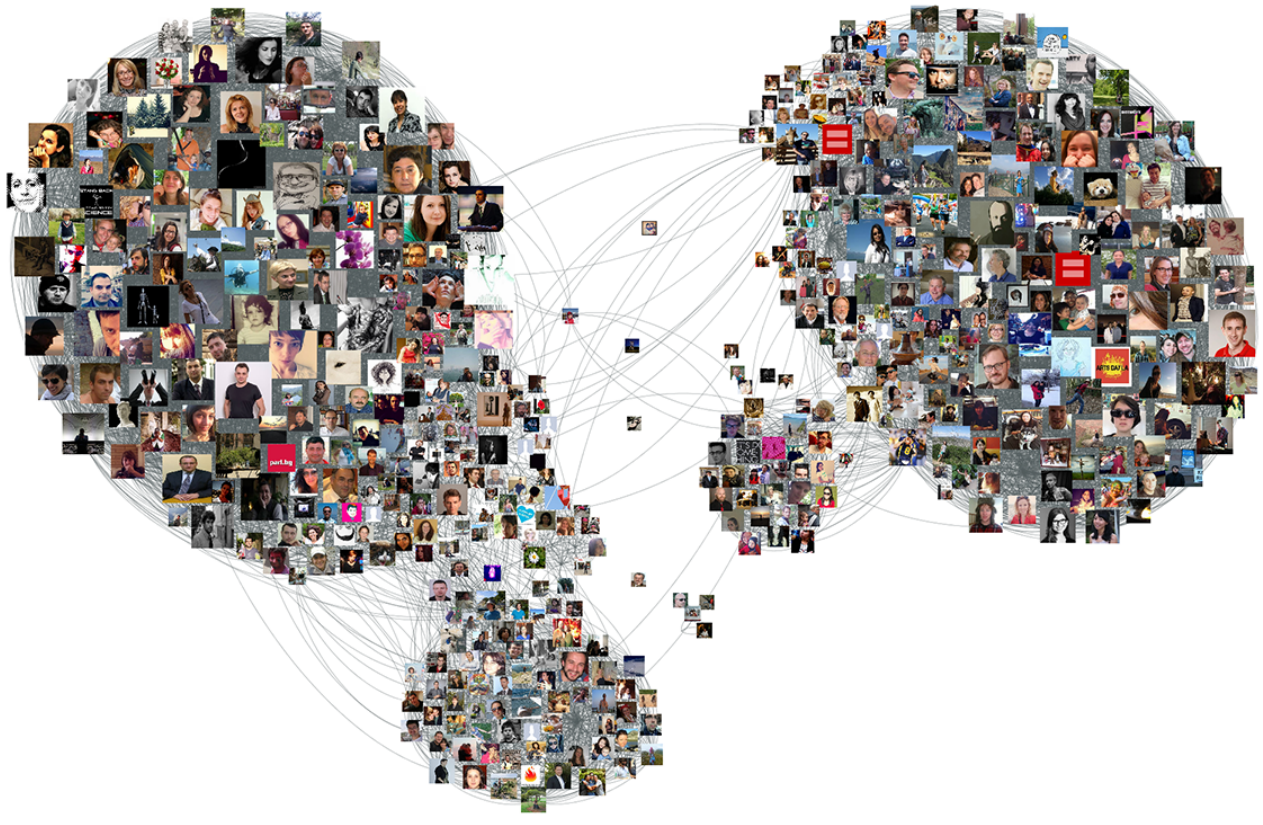
Han Zhang

Outline

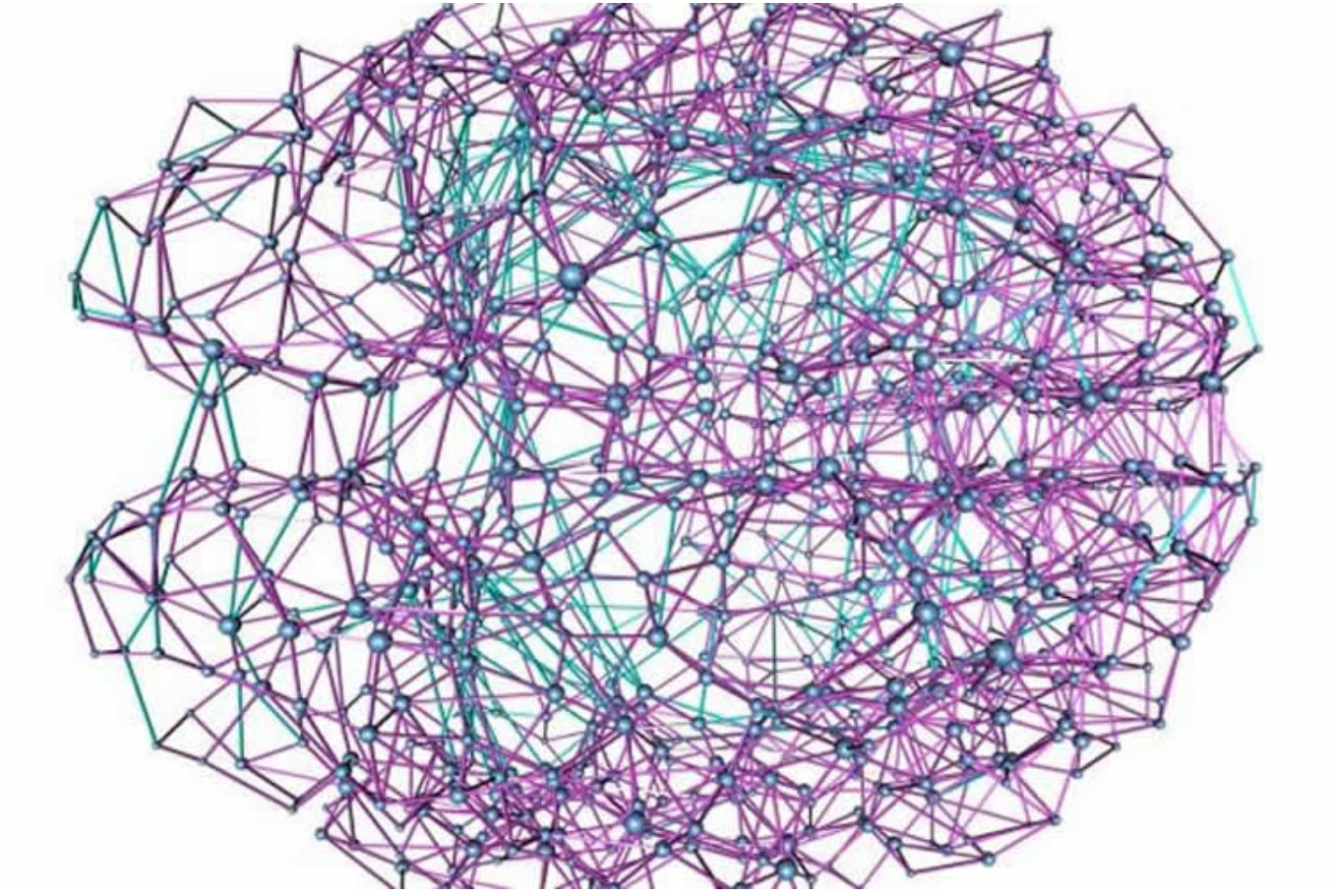
Captureing inter-connections

- Network analysis is a general theoretical approach to study connections
- Connections between
  - Humans
  - Cities
  - Facilities (e.g. airports)
- David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne, *Computational Social Science*, Science **323** (2009), no. 5915, 721–723

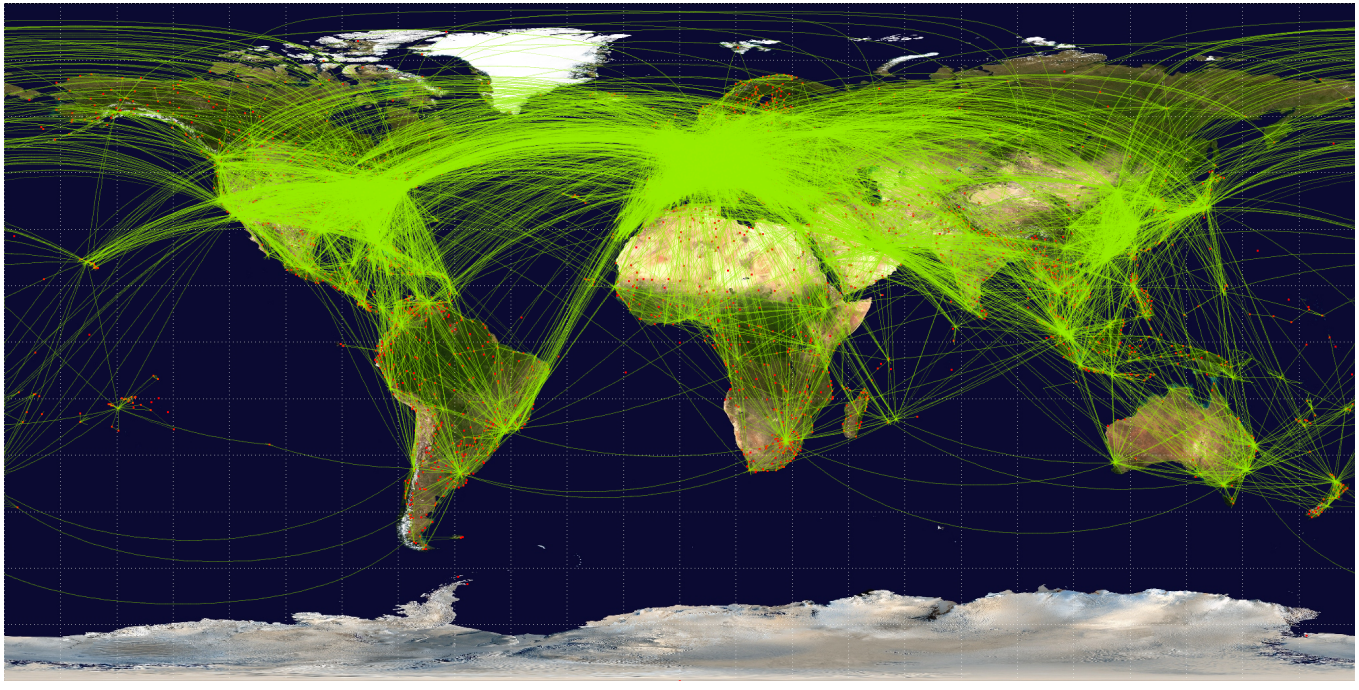
Facebook network



Brain neural networks



Global Aviation Network



Notations

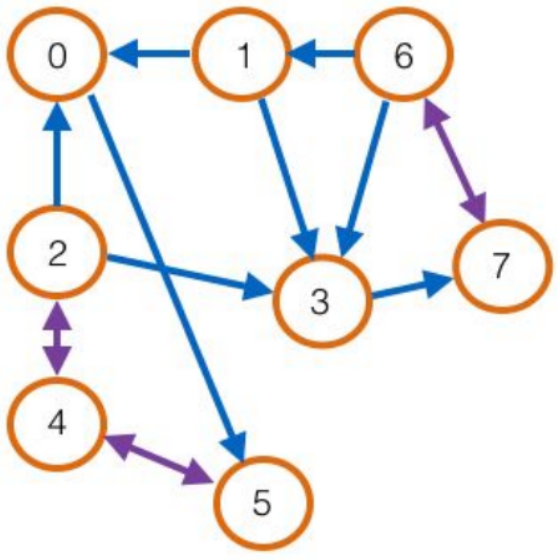
- Different networks can be abstracted in the following way
- Node or vertex: a unit in the network
  - e.g., a person, an airport, a neuron
- Edges: connection bewteen two units
  - Directed
  - Undirected
  - Self-loop
- Network/graph: nodes and edges together
  - $G =< V, E >$

Network Data

- A network can be stored in two ways
- G can be stored as an adjacency matrix A
  - If i are connected with j, then $A_{ij} = 1$; otherwise it's 0
  - In weighted network, $A_{ij} = 1$ is the weights
    - e.g., number of flights between two airports
- Or, G can be stored as an edge list or adjacency list
  - Especially when the network is sparse
  - Usually social networks are sparse

Edge list

Graph Representation: Edge List



From	To
0	5
1	3
1	0
2	0
2	3
2	4
3	7
4	2
4	5
5	4
6	1
6	3
6	7
7	6

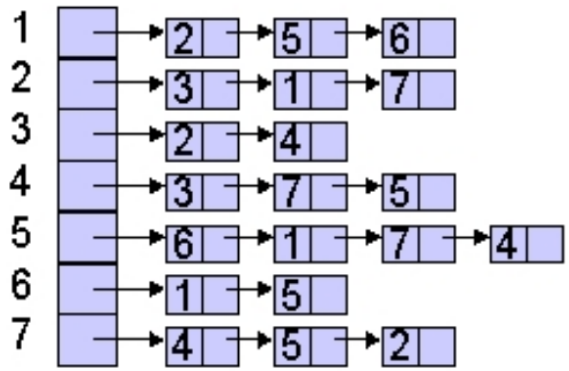
Edge list and adjacency list

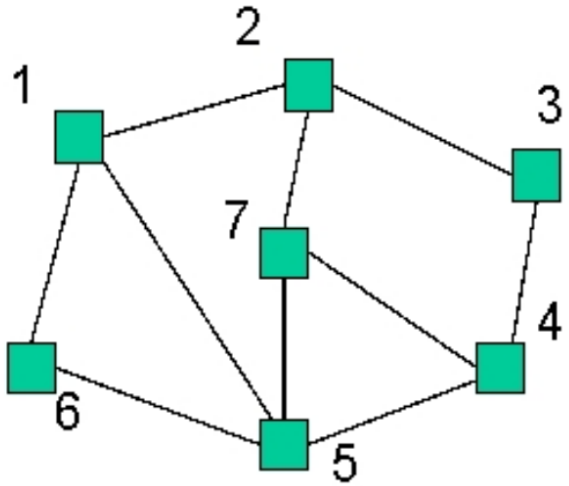
Edge List and Adjacency Lists

- List of edges

1	5	1	2	2	3	5	7	5	5
2	1	6	7	3	4	6	4	7	4

- Adjacency lists





Concepts

- Degree of node i, d_i: number of friends of a node i
- Degree characterizes the importance of a node in a network

Measurement of Network: ego-centered networks

- Ego-centered network: connections of a particular person
- The simpliest approach commonly used in surveys, using a techniques called name generator
  - That is, for each respondent, ask them to list five (or ten) of their best friends
  - And then let the respondent to answer some questions about their friends
  - Problems:
    - Recall biases
    - do not measure whether two friends of a respondent are friends
- Slightly more complicated version:
  - Ask whether two friends know each other

Ego-centered network: example

- GSS (General Social Survey)
- `https://gssdataexplorer.norc.org/variables/848/vshow`

  From time to time, most people discuss important matters with other people. Looking back over the last six months - who are the people with whom you discussed matters important to you? Just tell me their first names or initials. If LESS THAN 5 NAMES MENTIONED, PROBE, Anyone else? ONLY RECORD FIRST 5 NAMES.

- Why don't they directly ask "who is your friend?"

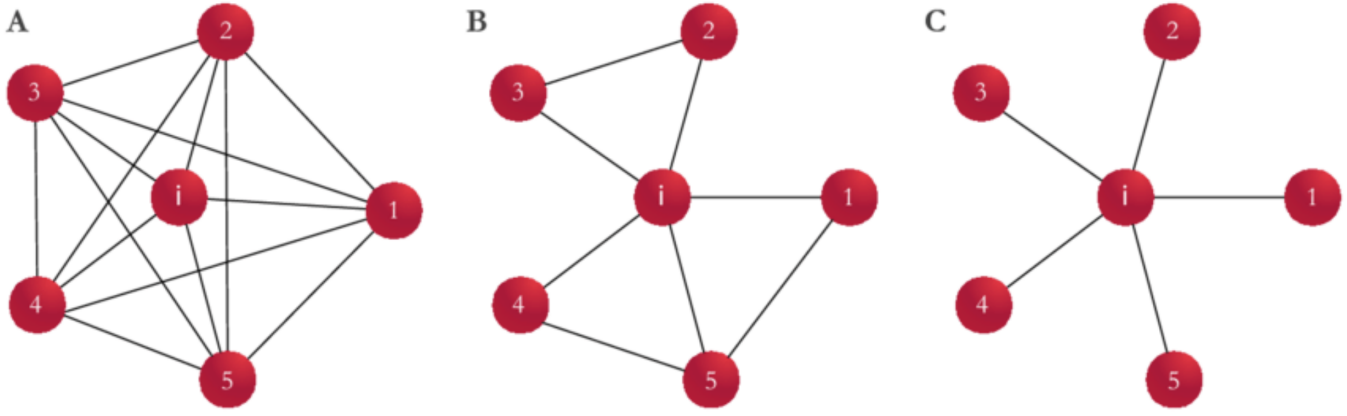
Ego-centered networks and loneliness

- Miller McPherson, Lynn Smith-Lovin, and Matthew E. Brashears, *Social Isolation in America: Changes in Core Discussion Networks over Two Decades*, American Sociological Review **71** (2006), no. 3, 353–375
- From 1985 to 2004, the number of listed discussants shrunk from 3 to 2. Americans became lonelier
- Claude S. Fischer, *The 2004 GSS Finding of Shrunken Social Networks: An Artifact?*, American Sociological Review **74** (2009), no. 4, 657–669
- It's just an interviewer effect: the question in 2004 survey appeared at the end so people were tired and did not want to answer

Clustering coefficient

- Beyond degree, human social network also exhibit a unique phemonena:
  - Nodes tend to create tightly knit groups characterised by a relatively high density of ties
  - In other words, one is more likely to befriend his friends of friends than to befriend a random person
- Local clustering coefficient characterizes this phenomena
- For each node i, local clustering coefficient $C_i = \frac{\text{Number of observed edges}}{\text{Number of possible edges}}$
- The numerator, number of possible edges for a node with d edges is $\frac{d*(d-1)}{2}$

Clustering coefficient



Measurement of Network: whole network

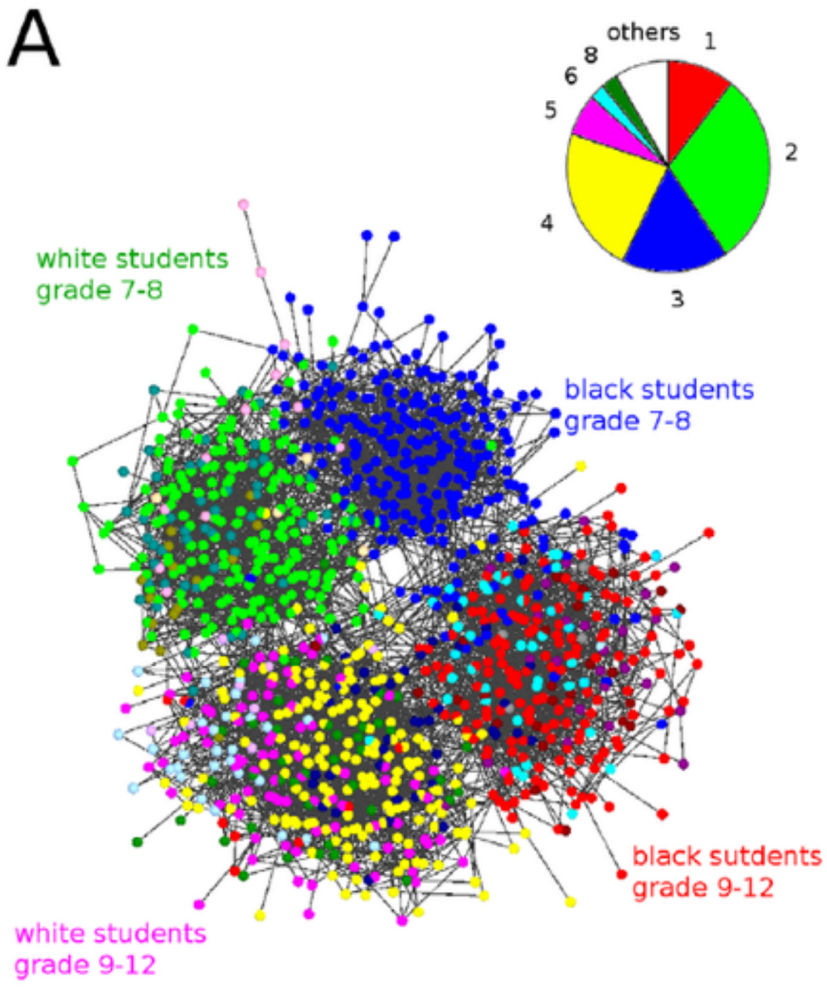
- Whole network: try to measure complete set of relationship between nodes
- Ideal measure, but very challenging
- Why? the connections you need to measure in quickly grows with the number of nodes N
  - you need to as on a scale of N^2, which could be crazy even if you are studying a small community such as all people in HKUST with 20,000 people
  - And people may not want to give their friends list, or they simply do not remember it
- Using surveys to measure whole network has only been successful in small and closed communities, such as schools or small villages

Community structure with whole network

- With whole network, we can explore community structure
- Basically unsupervised clustering of nodes into groups, with nodes insides a group more densely connected
- This provide information on social groups, not defined by their attributes but by their connections

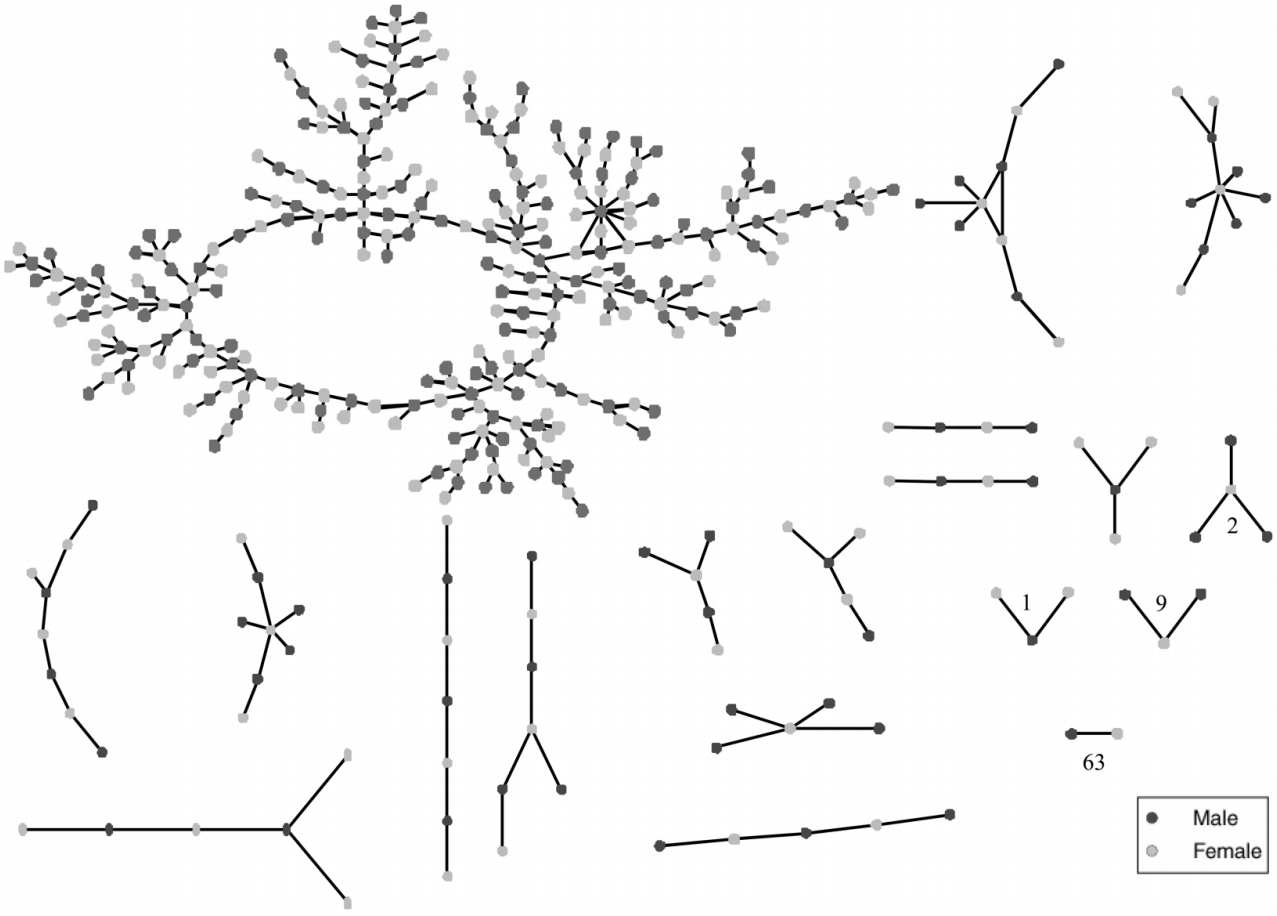
Add Health Friendship

- Add Health is a famous survey dataset that sampled 80 US schools and construct whole network within each school



Add Health Dating Network

- Peter S. Bearman, James Moody, and Katherine Stovel, *Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks*, American Journal of Sociology **110** (2004), no. 1, 44–91



Measuring whole networks in the age of big data

- With the rise of social networking sites and mobile phones, it is possible to measure whole network at large scale
  - We will look at an example shortly
  - But remember these are re-purposed data. Do Facebook capture all my friendships? Certainly not

Small-world experiment

- People found out how to study a particular aspect of the whole network structure—the distance between two nodes—without actually measuring the whole network
- It's known as the "small-world phenomena," or the "six degree of separation"

  I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation between us and everyone else on this planet. The President of the United States, a gondolier in Venice, just fill in the name. . . It's not just the big names. It's everyone. A native in the rain forest. An Eskimo. I am bound to everyone on this planet by a trail of six people. It's a profound though.

- `https://www.imdb.com/title/tt0108149/?ref_=tt_ch`

History of Small-World Experiment

- Let us go back to 1967
- `http://snap.stanford.edu/class/cs224w-readings/milgram67smallworld.pdf`

An Experimental Study of the Small World Problem*

JEFFREY TRAVERS

Harvard University

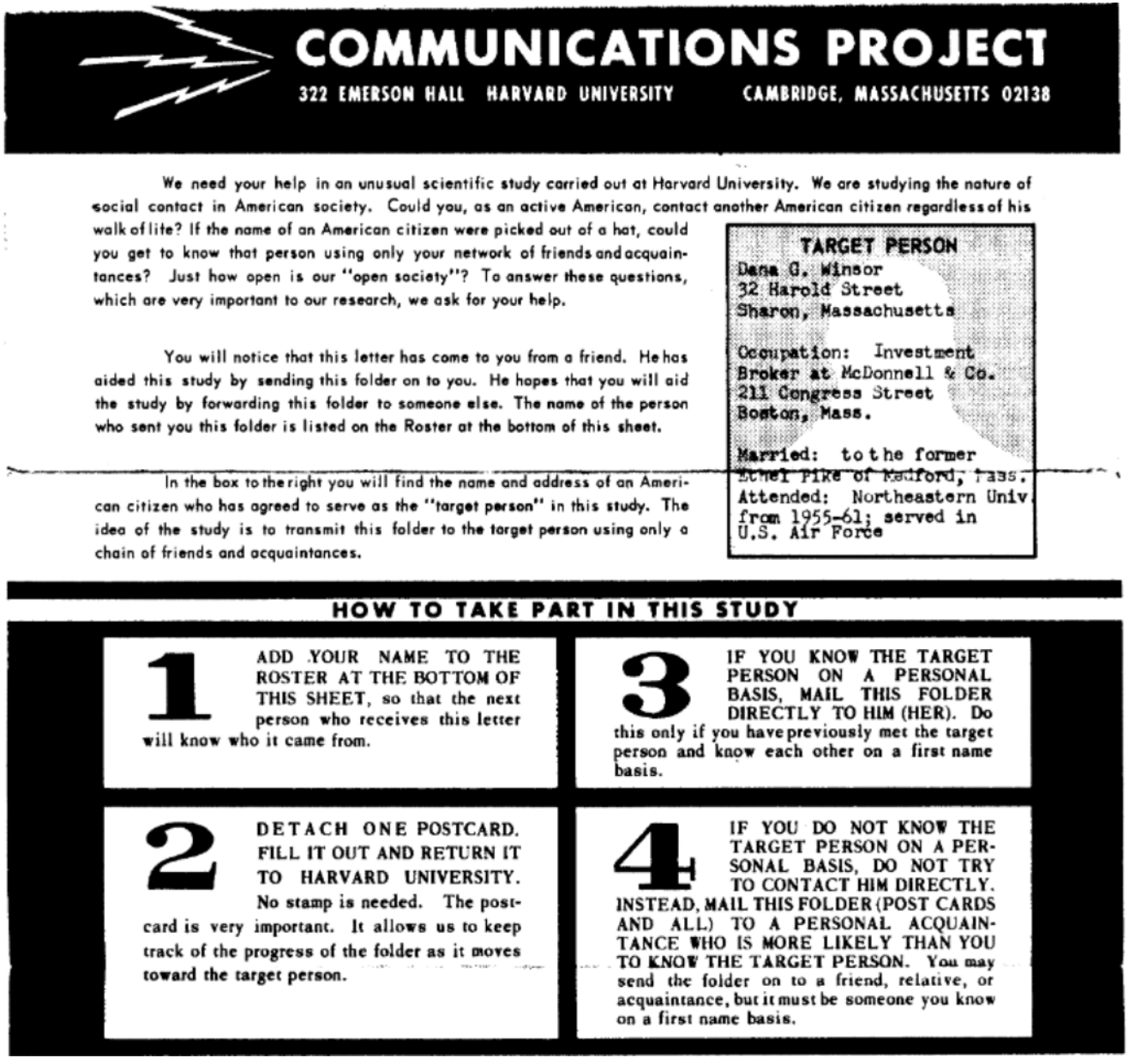
AND

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.

Questionnaire used



COMMUNICATIONS PROJECT

322 EMERSON HALL HARVARD UNIVERSITY CAMBRIDGE, MASSACHUSETTS 02138

We need your help in an unusual scientific study carried out at Harvard University. We are studying the nature of social contact in American society. Could you, as an active American, contact another American citizen regardless of his walk of life? If the name of an American citizen were picked out of a hat, could you get to know that person using only your network of friends and acquaintances? Just how open is our "open society"? To answer these questions, which are very important to our research, we ask for your help.

You will notice that this letter has come to you from a friend. He has aided this study by sending this folder on to you. He hopes that you will aid the study by forwarding this folder to someone else. The name of the person who sent you this folder is listed on the Roster at the bottom of this sheet.

In the box to the right you will find the name and address of an American citizen who has agreed to serve as the "target person" in this study. The idea of the study is to transmit this folder to the target person using only a chain of friends and acquaintances.

TARGET PERSON

Dana G. Winsor
32 Harold Street
Sharon, Massachusetts

Occupation: Investment Broker at McDonnell & Co.
211 Congress Street
Boston, Mass.

Married: to the former Ethel Pike of Medford, Mass.
Attended: Northeastern Univ. from 1955-61; served in U.S. Air Force

HOW TO TAKE PART IN THIS STUDY

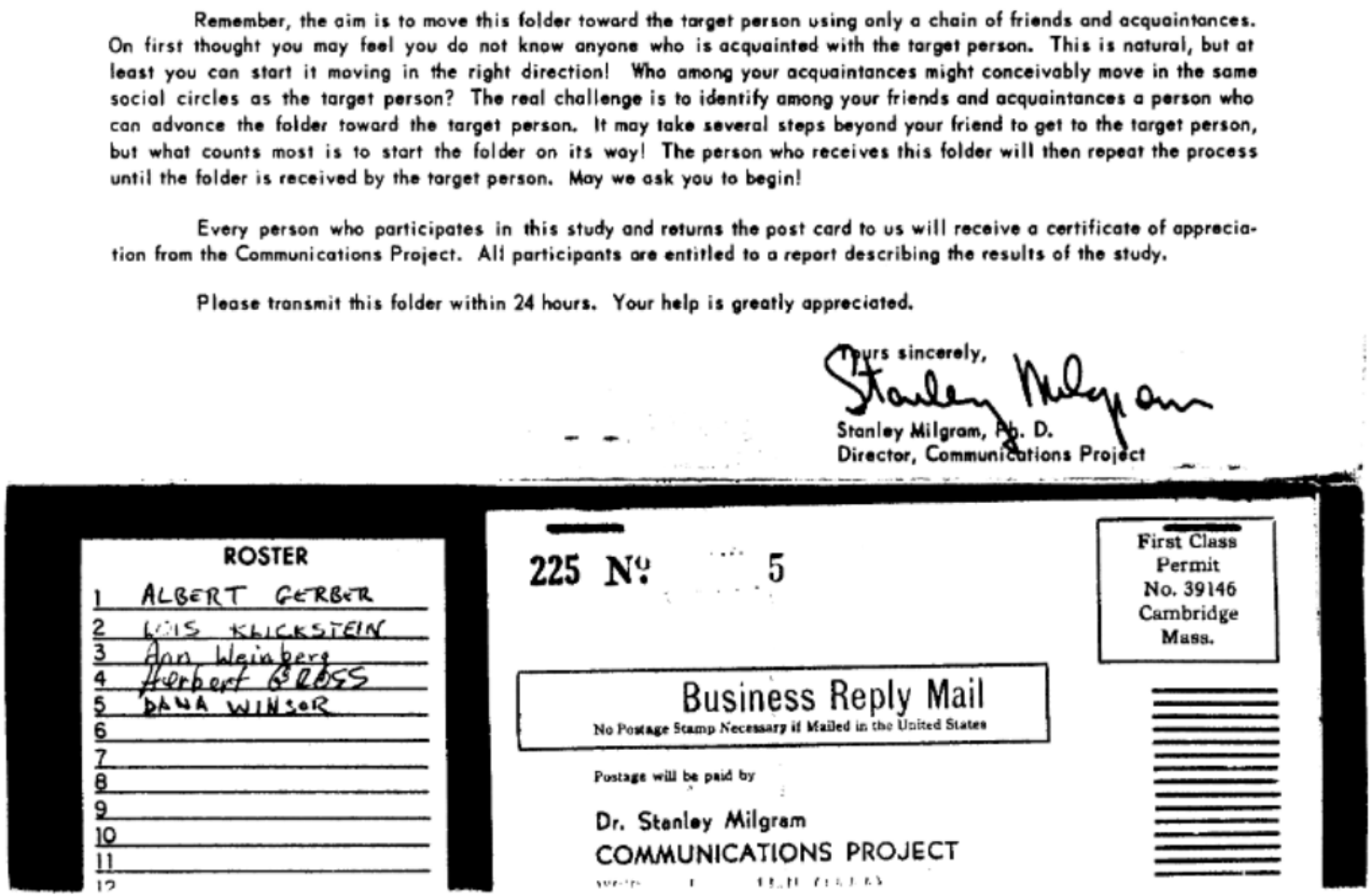
1 ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THIS SHEET, so that the next person who receives this letter will know who it came from.

2 DETACH ONE POSTCARD. 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.

3 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.

4 IF YOU DO NOT KNOW THE TARGET PERSON ON A PERSONAL BASIS, DO NOT TRY TO CONTACT HIM DIRECTLY. INSTEAD, MAIL THIS FOLDER (POST CARDS AND ALL) TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON. You may send the folder on to a friend, relative, or acquaintance, but it must be someone you know on a first name basis.

Questionnaire used



Remember, the aim is to move this folder toward the target person using only a chain of friends and acquaintances. On first thought you may feel you do not know anyone who is acquainted with the target person. This is natural, but at least you can start it moving in the right direction! Who among your acquaintances might conceivably move in the same social circles as the target person? The real challenge is to identify among your friends and acquaintances a person who can advance the folder toward the target person. It may take several steps beyond your friend to get to the target person, but what counts most is to start the folder on its way! The person who receives this folder will then repeat the process until the folder is received by the target person. May we ask you to begin!

Every person who participates in this study and returns the post card to us will receive a certificate of appreciation from the Communications Project. All participants are entitled to a report describing the results of the study.

Please transmit this folder within 24 hours. Your help is greatly appreciated.

Yours sincerely,

Stanley Milgram, Ph. D.
Director, Communications Project

ROSTER

1 ALBERT GERBER
2 LOIS KLICKSTEIN
3 Ann Weinberg
4 Herbert GROSS
5 DANA WINSOR
6
7
8
9
10
11
12

225 No. 5

First Class Permit No. 39146 Cambridge Mass.

Business Reply Mail
No Postage Stamp Necessary if Mailed in the United States

Postage will be paid by

Dr. Stanley Milgram
COMMUNICATIONS PROJECT

- https://en.wikipedia.org/wiki/Small-world_experiment#/media/File:Experement_Small_World_(possible_option).gif

Result 1

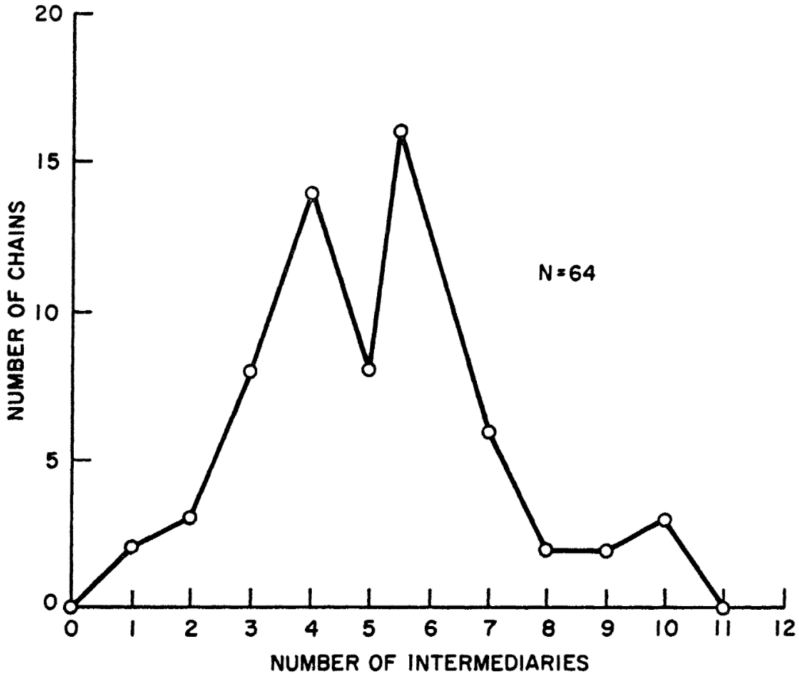


FIGURE 1

Lengths of Completed Chains

- The mean number of intermediaries: 5.2
- In other words, six degress of separation to reach a random person in the US

Result 2

- 29% of chains reached targets.
- [In class discussion]: What are the implications?

Result 2

- What does this tells you?

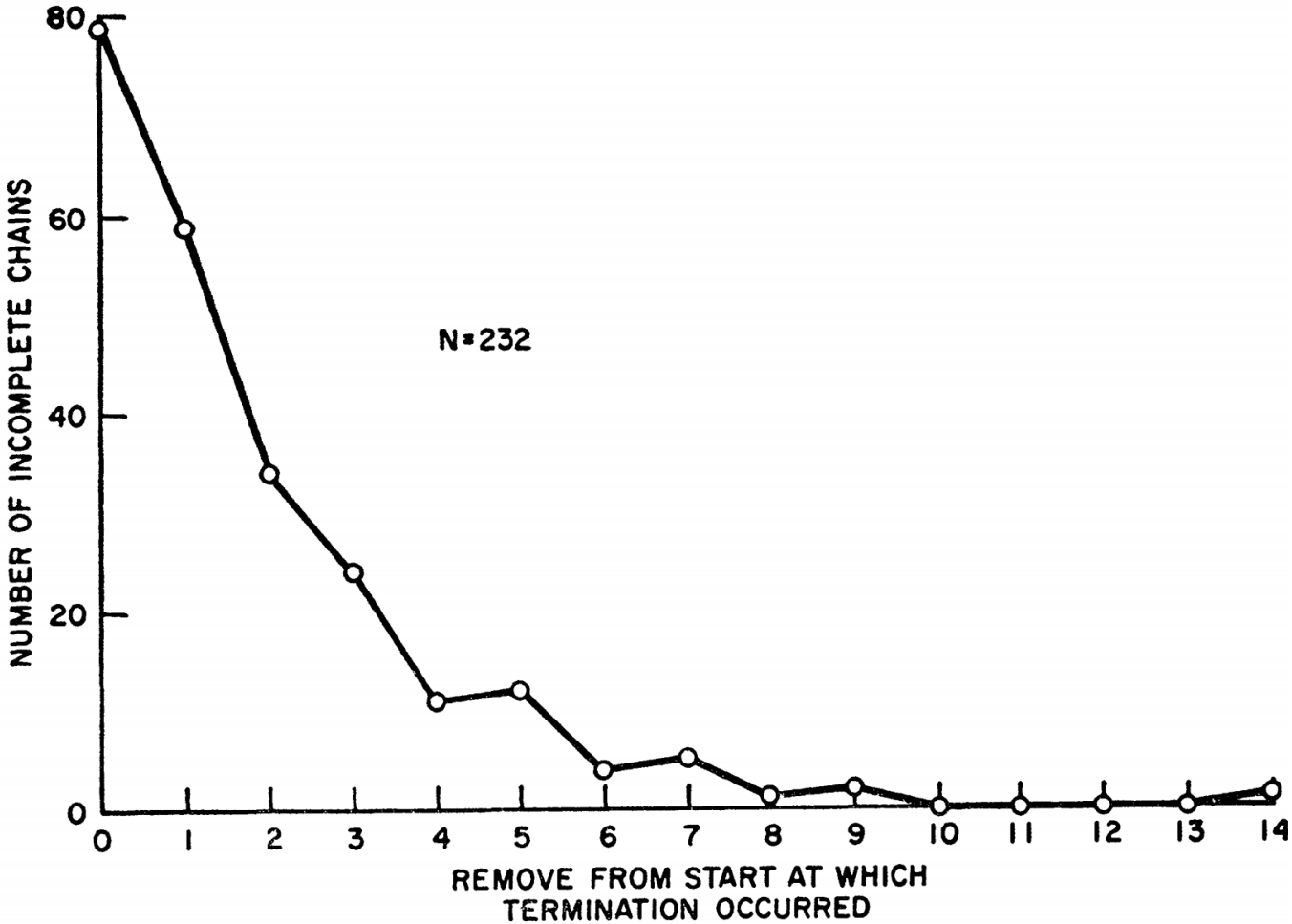


FIGURE 2

Lengths of Incomplete Chains

Result 3

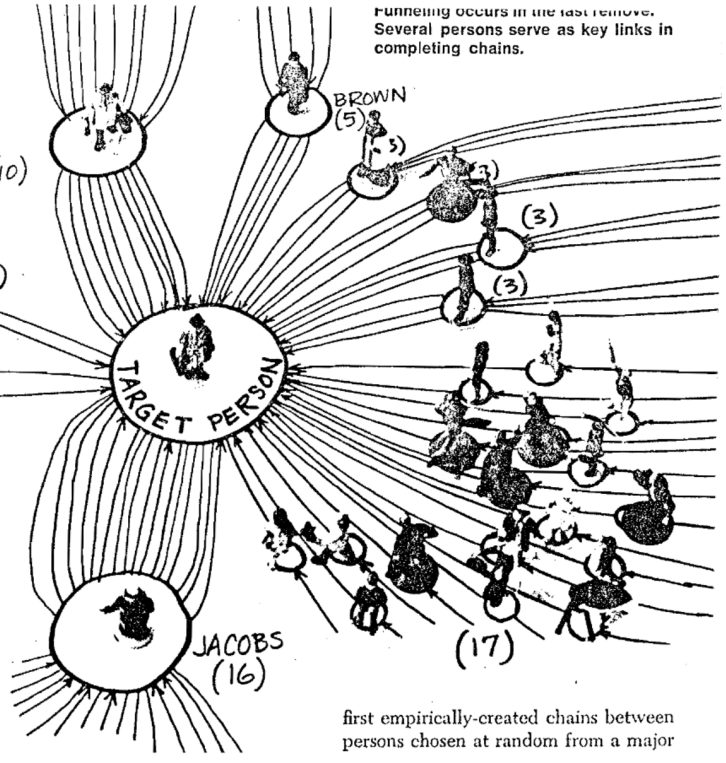
- The 64 letters which reached the target were sent by a total of 26 people.
- Sixteen, fully 25 per cent, reached the target through a single neighbor.
- Another 10 made contact through a single business associate
- And 5 through a second business associate.
- These three superconnectors together accounted for 48% of the total completions

Result 3

arch extends to an enor-
f persons.
ı we state there are only
e acquaintances, this
ness between the position of
son and the target person.
ge measure
nfusion of
ferent frames
two persons are
art, they
leed. Almost
nited States is
ves from
r from
ller,

if a
ematical
does not,
l sense,
ves with that of Nelson
us, when we speak of
ies, we are talking about
ychological distance be-
ng and target points, a
eems small only because
regard "five" as a small
ntity. We should think
ts as being not five per-
"five circles of acquaint-
five "structures" apart.
et it in its proper per-

Funneling occurs in the last remove. Several persons serve as key links in completing chains.



first empirically-created chains between persons chosen at random from a major

Modern replications of Milgram's Small World Experiment

- Peter Sheridan Dodds, Roby Muhamad, and Duncan J. Watts, *An Experimental Study of Search in Global Social Networks*, Science **301** (2003), no. 5634, 827–829

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.

Result 1

- Completion rate, any guess?
- "Although the average participation rate (about 37%) was high relative to those reported in most e-mailbased 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)."

Result 2

- The mean length is 4.05.
- Accounting for incompletion, the estimated median length is 5 for pairs in the same country and 7 for pairs in different countries

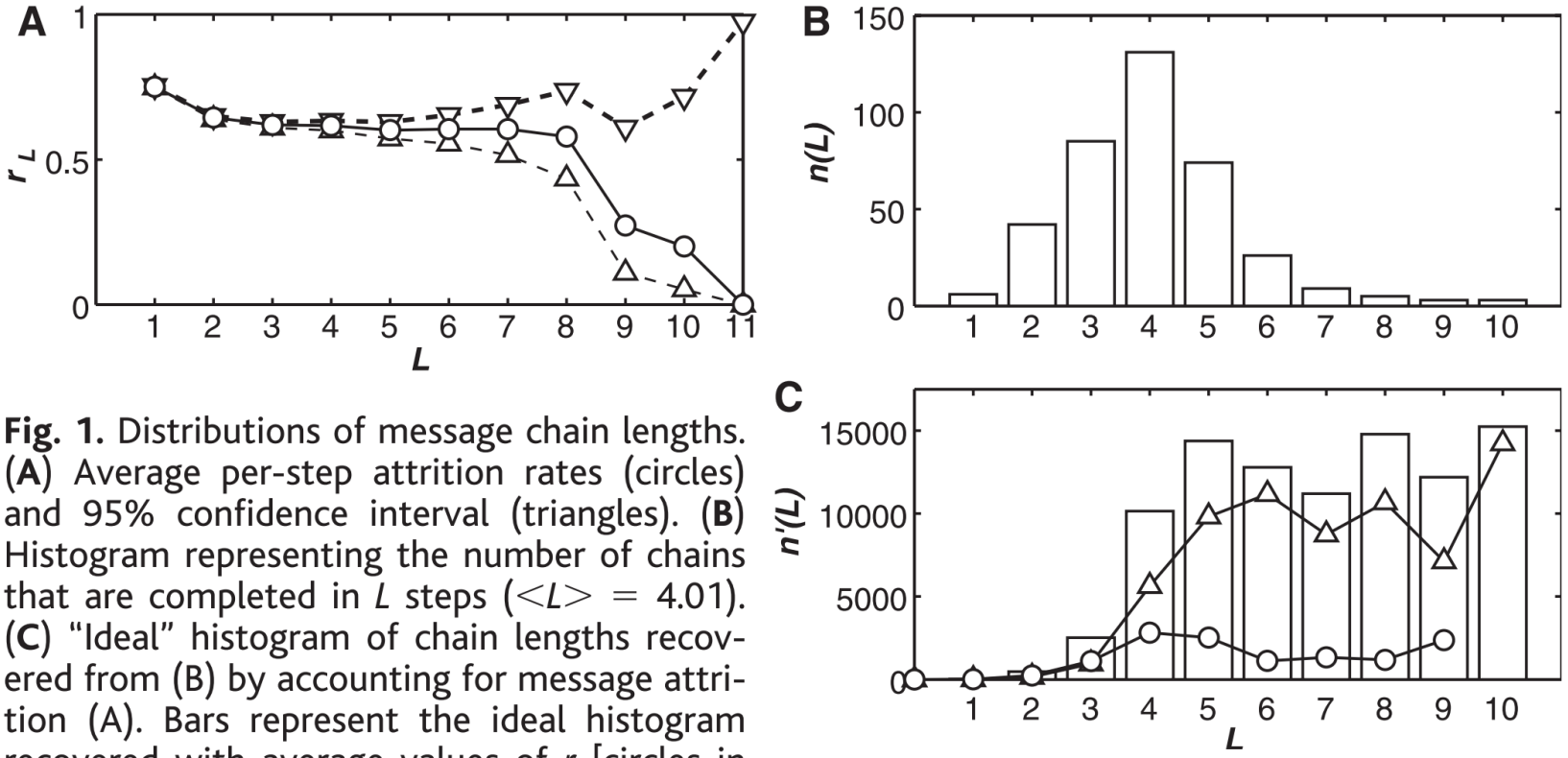


Fig. 1. Distributions of message chain lengths. (**A**) Average per-step attrition rates (circles) and 95% confidence interval (triangles). (**B**) Histogram representing the number of chains that are completed in L steps ($<L> = 4.01$). (**C**) "Ideal" histogram of chain lengths recovered from (B) by 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

Result 3

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

Measuring degree of separation with social media

- `https://research.fb.com/three-and-a-half-degrees-of-separation/`

Year	Distance	Number of Facebook users
2008	5.28	5.8 million users
2011	4.74	721 million users
2016	4.57	1.6 billion users

Degree of separation of other objects

- Average distance between random web pages. Guess?
  - It's 19 in 1999
  - Réka Albert, Hawoong Jeong, and Albert-László Barabási, *Diameter of the World-Wide Web*, Nature **401** (1999), no. 6749, 130–131
- Average distance between two Wikipedia pages.
  - Let us play some games
  - Write down your path from two words

Degree of separation on Wikipedia

- `https://degreesofwikipedia.com/`
- Is your result similar to the true value?
- "Individuals operating with purely local information are very adept at finding these chains. "
  - Jon M. Kleinberg, *Navigation in a small world*, Nature **406** (2000), no. 6798, 845–845

Small-world network definitions

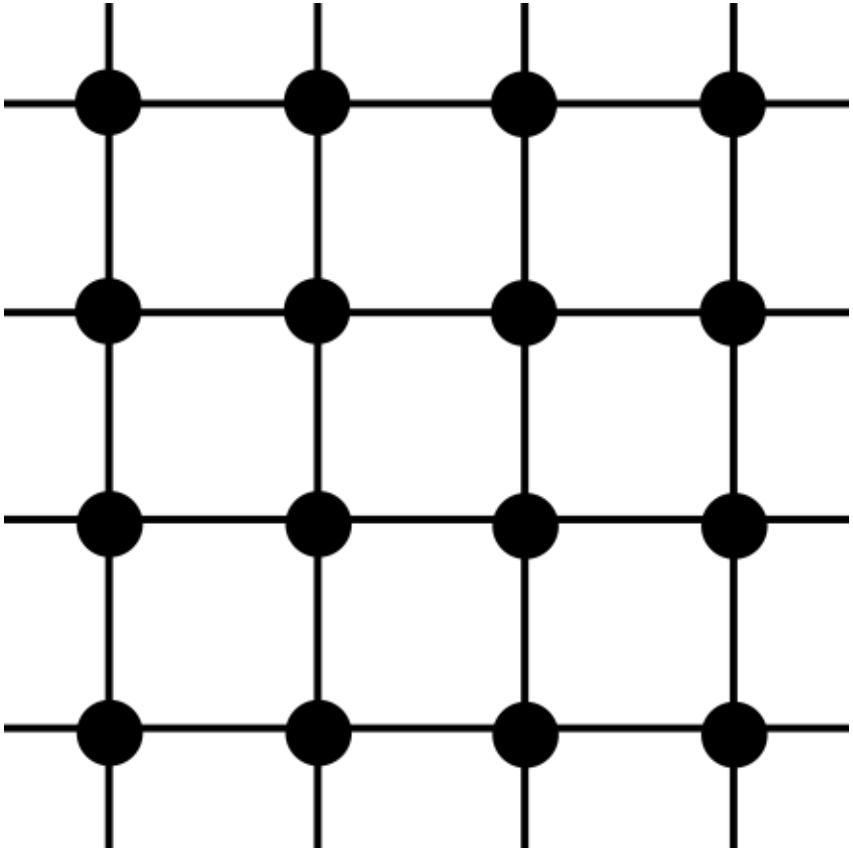
- A small-world network needs to have:
  - Average shortest distance between any pairs of nodes (diameter) is on a scale of $log(N)$, where N is the number of nodes;
    - The diameter is thus much smaller than the network size
  - Clustering coefficient is not too small
    - i.e., you should have higher probability to befriend with your friends of friends, than some random person

Theoretical modeling of small-world networks

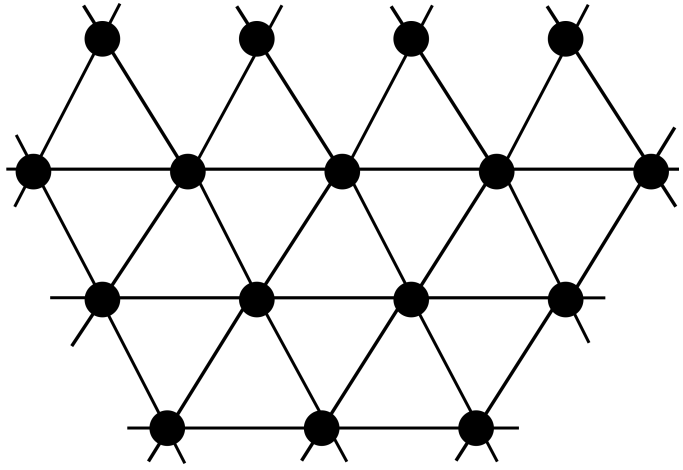
- So far, we have taken an empirical approach toward understanding small-world networks
  - e.g., measuring/describing the diameter
- A theoretical approach, however, asks what's the underlying conditions that produce small-world networks
- Before answering these questions, we first look at two simple network examples that are not small-world networks

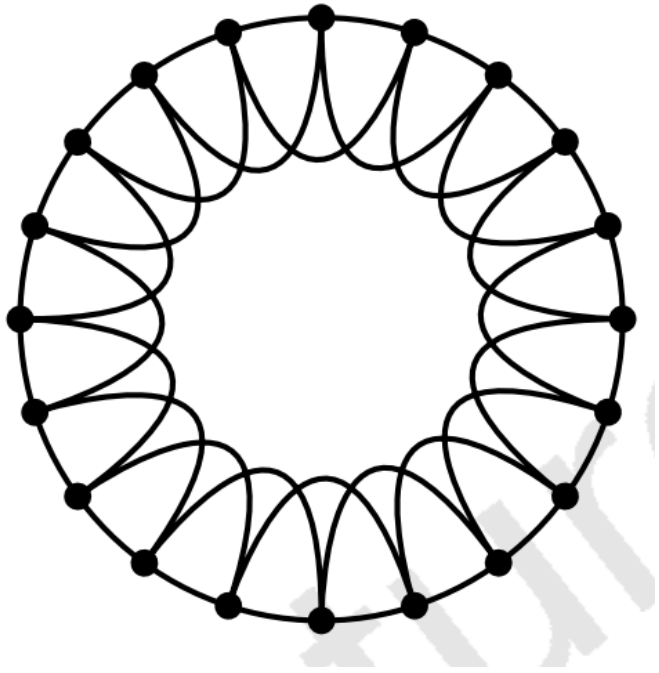
Simplest network: regular network

- Every node has the exact same number of edges



Simplest network: regular network



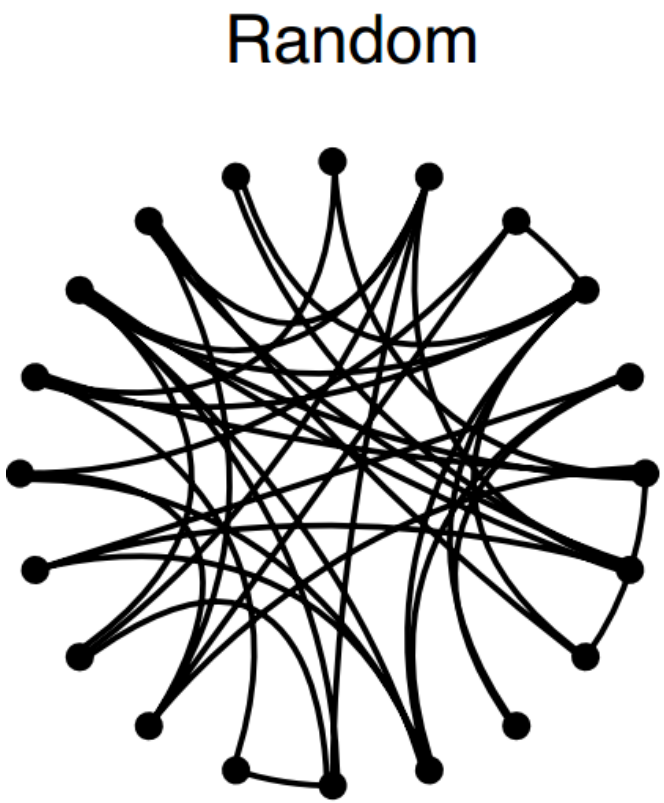


Regular networks are not small-world networks

- Why?
- Diameter is large

Simplest network model: Erdos-Renyi network

- Erdos-Renyi network: the simplest random network
- N nodes
- Each node has a probability to make an edge with any other with a probability p
- $N \cdot p$ edges in expectation



Erdos-Renyi network is not small-world network

- `http://www.networkpages.nl/CustomMedia/Animations/RandomGraph/ERRG/AddoneEdgepATime.html`
- Is this network small world? No
- The diameter is small enough
- But the clustering coefficient $\rightarrow 0$ when N increases

Comparisons

- The two ideal types, regular networks and random networks, looks very different
- And they are all different from small-world networks

	Diameter L	Clustering Coefficient C
Random	small	small
Regular	large	large
Small-world	small, around $log(N)$	large

Small-world Phenomena beyond social networks

- "the small-world phenomenon is not merely a curiosity of social networks, nor an artefact of an idealized modelit is probably generic for many large, sparse networks found in nature"

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.25	0.28	0.05

Diameter vs. Clustering Coefficients

- More on diameter vs. clustering coefficients
- Diameter is a global measure
  - It's the average shortest distance between each pairs of nodes
- Local clustering coefficient is a local measure
  - You can collect more complete information about an indidividual, by asking whether two of his friends know each other
- Local clustering coefficient is easier to measure than diameter

Watts-Strogatz Model

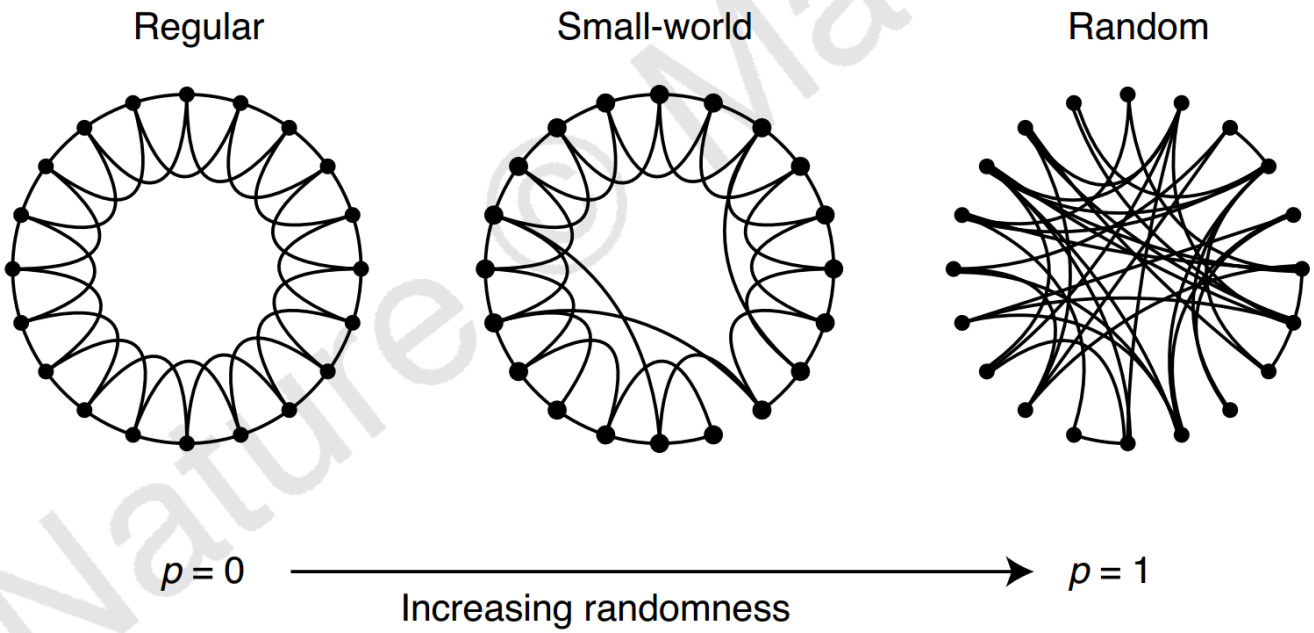
- Duncan J. Watts and Steven H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature **393** (1998), no. 6684, 440–442
- Perhaps the most influential work of modern network analyis
- Key intuition: only adding several long-range edges can turn regular network into a small-world network
- Why? These long-range edges connect otherwise distance nodes

Watts-Strogatz Model

- Theory building from simulation, or agent-based modeling
  - Start from a regular network
  - "We choose a vertex and the edge that connects it to its nearest neighbour in a clockwise sense. With probability p, we reconnect this edge to a vertex chosen uniformly at random over the entire ring, with duplicate edges forbidden; otherwise we leave the edge in place."
  - Increasing p makes the graph more random
  - $p = 1$ makes the network completely random
- Demo: `http://www.netlogoweb.org/launch`

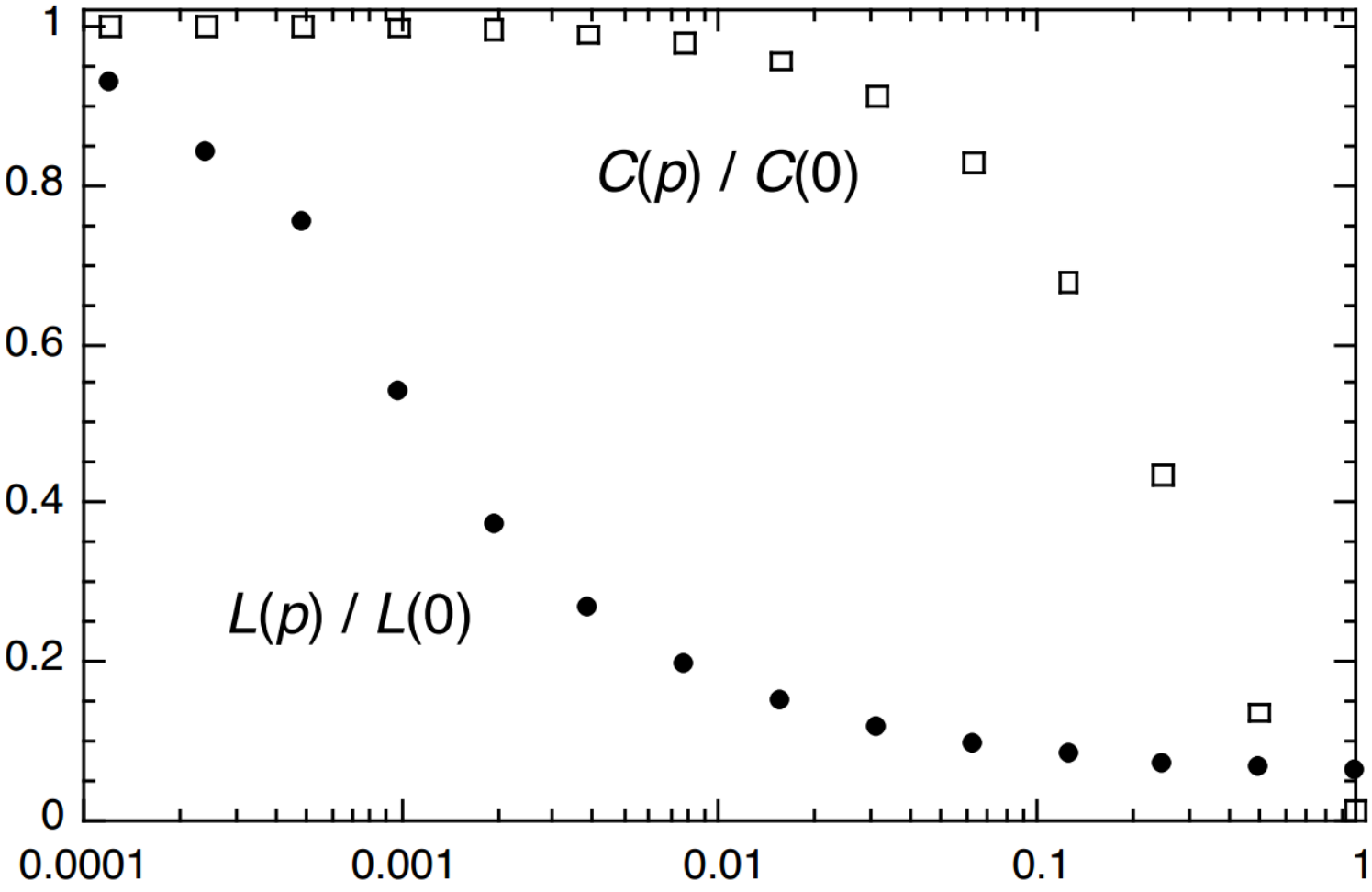
Watts-Strogatz Model

- Key finding from Watts and Strogatz:
  - a very small number of p would suffice to turn a regular network into small-world network



Watts-Strogatz Model

- A small p leads to a small-world network
  - Large clustering coefficient $C(p)$
  - Small diameter $L(p)$



Two types of computational social sciences

- Two parallel developments of computational social sciences
- For studying complex networks
  - Social phenomena are non-linear; we need to study it as a complex network
  - A natural hybrid of theory-driven mathematical simulations and empirical analysis using big data
  - A new paradigm; from studying attributes to studying connections; big mind shift.
- For measurement
  - E.g., applying machine learning techniques on text data to generate some variables, and then put these variables into a linear regressions to test some theories
  - Mostly an empirical approach: old theory + new data