Network Analysis, softer stuff

CIS 600, Spring 2018

February 1, 2018

Today - History & Applications

- Small Worlds
- Influence
- Community

The Small World Phenomenon

- In pop culture:
  - Refers to the idea that the world looks small when you think of how short a path of friends it takes to get from you to almost anyone else.

  - Hungarian writer, Frigyes Karinthy, in his 1929 short story "Láncszemek" ("Chains") suggested that any two persons are distanced by at most six friendship links.

  - Better known as the six degrees of separation, a phrase coming from the play of this title by John Guare (1990).

    "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 Milgram Experiment

- The first experimental study, and the origin of the number **six** in the pop-culture, was performed by Stanley Milgram and his colleagues in the 1960s (*Psych Today* **2**, 60,1967)

- Lacking massive social-network datasets we have today, and with a budget of only $680, he set out to test the idea that people are really connected in the global friendship network by short chains of friends.

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages

Stanley Milgram

From Wikipedia, the free encyclopedia

Stanley Milgram (August 15, 1933 – December 20, 1984) was an American social psychologist, best known for his controversial experiment on obedience conducted in the 1960s during his professorship at Yale.[4] Milgram was influenced by the events of the Holocaust, especially the trial of Adolf Eichmann, in developing the experiment.

After earning a Ph.D. in social psychology from Harvard University, he taught at Yale, Harvard, and then for most of his career as a professor at the City University of New York Graduate Center, until he died in 1984. His small-world experiment while at Harvard led researchers to analyze the degree of connectedness, including the six degrees of separation concept. Later in his career, Milgram developed a technique for creating interactive hybrid social agents (cyranoids), which has since been used to explore aspects of social- and self-perception.

He is widely regarded as one of the most important figures in the history of social psychology. A *Review of General Psychology* survey, published in 2002, ranked Milgram as the 46th-most-cited psychologist of the 20th century.[5]

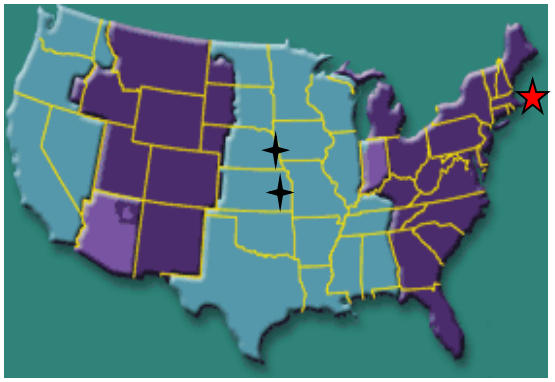
Contents [hide]

Stanley Milgram

Born	August 15, 1933 New York City, New York, U.S.
Died	December 20, 1984 (aged 51) Manhattan, New York City, New York, U.S.
Cause of death	Heart attack[1]
Education	Queens College, New York (B.A., Political Science, 1954) Harvard University (Ph.D., Social Psychology, 1960)
Known for	Milgram experiment Small world experiment Familiar stranger
Title	Professor[2]
Spouse(s)	Alexandra Menkin Milgram

The Milgram Experiment

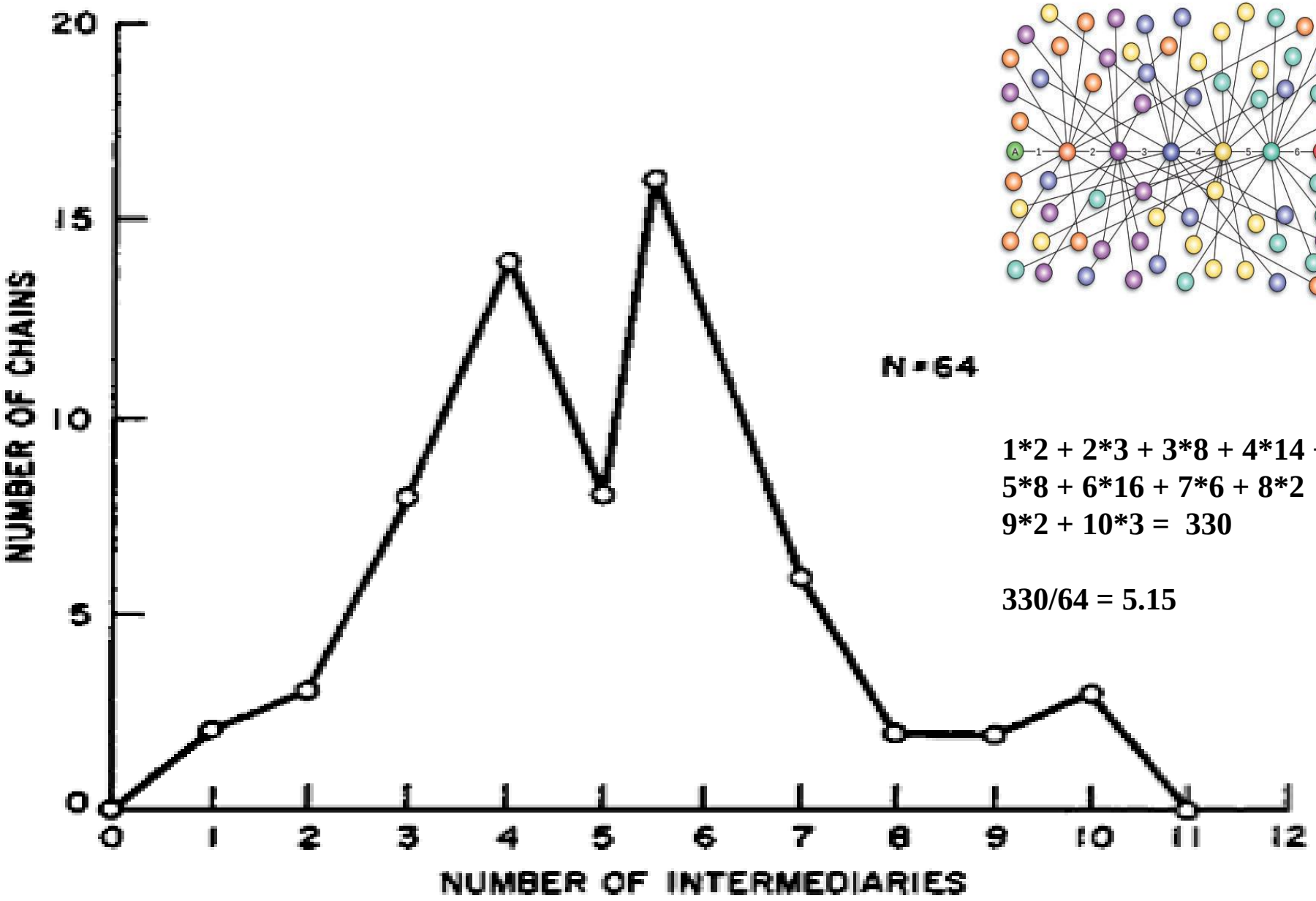
- He asked a collection of **296** randomly chosen starters (in Wichita, KS & Omaha, NE) to try forwarding a letter to a target person, a stockbroker who lived in a suburb of Boston (Sharon, MA).



- The starters were each given some personal info about the target, including address & occupation, and were asked to forward the letter to someone they knew on a first-name basis, with the same instructions, in order to eventually reach the target as quickly as possible.

The Milgram Experiment

The Milgram Experimental Results

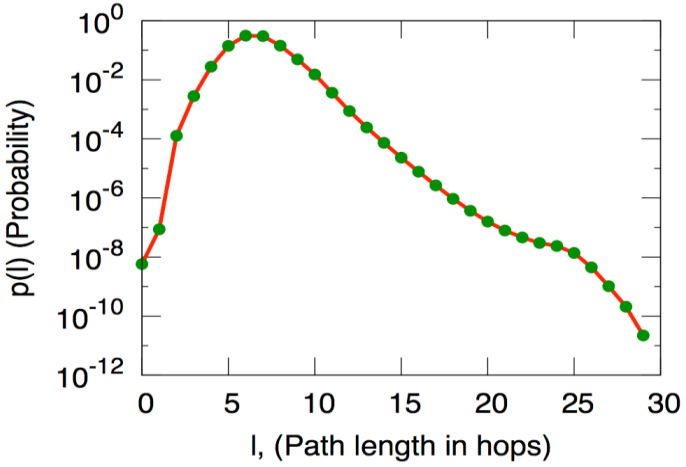


1*2 + 2*3 + 3*8 + 4*14 +
5*8 + 6*16 + 7*6 + 8*2
9*2 + 10*3 = 330

330/64 = 5.15

The Microsoft IM Experiment

- Jure Leskovec & Eric Horvitz (2008) analyzed the 240 million active user accounts on Microsoft Instant Messenger.
  - In the resulting graph, each node corresponds to a user, and there is an edge between two users if they engaged in a two-way conversation at any point during a month-long observation period.



  - As employees of Microsoft at the time, they had access to a complete snapshot of the system for the month under study - no missing data.
  - This graph turned out to have a giant component containing almost all of the nodes. And the average distance within this giant component were very small: **6.6**

J. Leskovec and E. Horvitz. Planetary-Scale Views on a Large Instant-Messaging Network, *Proceedings of WWW 2008*, Beijing, China, April 2008.

Planetary-Scale Views on a Large Instant-Messaging Network

Jure Leskovec*
Carnegie Mellon University
jure@cs.cmu.edu

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

ABSTRACT

We present a study of anonymized data capturing a month of high-level communication activities within the whole of the Microsoft Messenger instant-messaging system. We examine characteristics and patterns that emerge from the collective dynamics of large numbers of people, rather than the actions and characteristics of individuals. The dataset contains summary properties of 30 billion conversations among 240 million people. From the data, we construct a communication graph with 180 million nodes and 1.3 billion undirected edges, creating the largest social network constructed and analyzed to date. We report on multiple aspects of the dataset and synthesized graph. We find that the graph is well-connected and robust to node removal. We investigate on a planetary-scale the oft-cited report that people are separated by "six degrees of separation" and find that

We explore a dataset of 30 billion conversations generated by 240 million distinct users over one month. We found that approximately 90 million distinct Messenger accounts were accessed each day and that these users produced about 1 billion conversations, with approximately 7 billion exchanged messages per day. 180 million of the 240 million active accounts had at least one conversation on the observation period. We found that 99% of the conversations occurred between 2 people, and the rest with greater numbers of participants. To our knowledge, our investigation represents the largest and most comprehensive study to date of presence and communications in an IM system. A recent report [6] estimated that approximately 12 billion instant messages are sent each day. Given the estimate and the growth of IM, we estimate that we captured approximately half of the world's IM communication during the observation period.

The Erdos Numbers

- **Collaboration graph:** nodes correspond to mathematicians, and edges connect pairs who have jointly authored a paper with Paul Erdos, a mathematician who published roughly 1500 papers over his career.

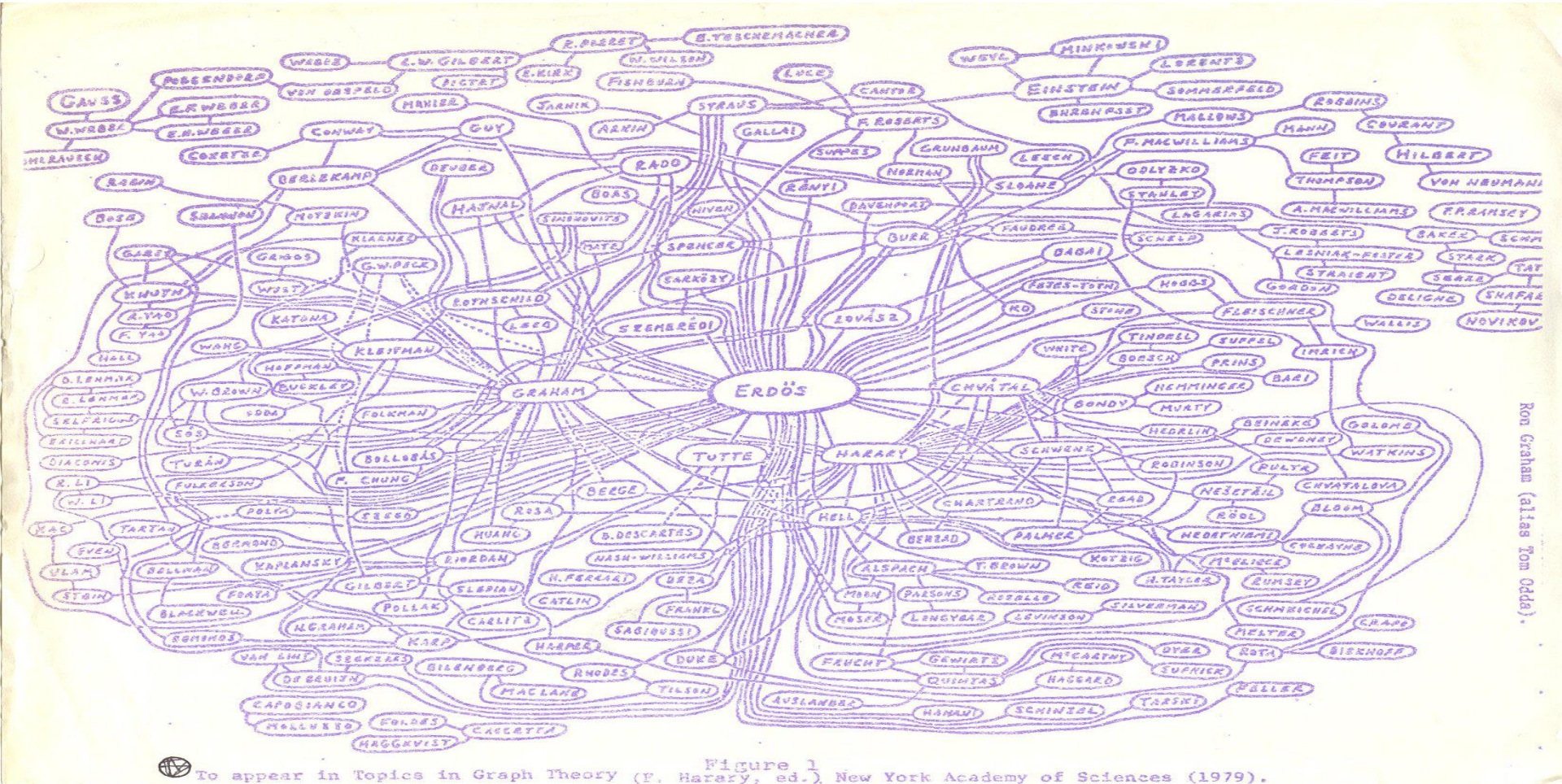
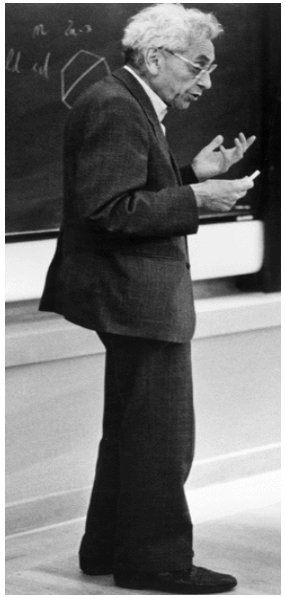


Figure 1

Ron Graham (alias Tom Odda).

The Erdos Numbers

- A mathematician's Erdos number is defined as the distance from him or her to Erdos in this graph
- Most **mathematicians** have Erdos numbers of at most 4 or 5
- Extending the collaboration graph to include co-authorship across **all sciences**, most scientists in other fields have Erdos numbers that are comparable or only slightly larger
  - Albert Einstein's is 2
  - Enrico Fermi's is 3
  - Noam Chomsky's is 4
  - Francis Crick’s is 5
  - James Watson's is 6
- The world of science is truly a small one in this sense.



WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Print/export

Languages
Alemannisch
العربية
Aragonés
Asturianu
Azərbaycanca
বাংলা
Bân-lâm-gú
Башкортса
Беларуская
Беларуская (тарашкевіца)
Български

Enrico Fermi

From Wikipedia, the free encyclopedia

"Fermi" redirects here. For other uses, see Fermi (disambiguation).

Enrico Fermi (Italian: [enˈri.ko ˈfer.mi]; 29 September 1901 – 28 November 1954) was an Italian physicist, best known for his work on Chicago Pile-1 (the first nuclear reactor), and for his contributions to the development of quantum theory, nuclear and particle physics, and statistical mechanics. He is one of the men referred to as the "father of the atomic bomb".[4] Fermi held several patents related to the use of nuclear power, and was awarded the 1938 Nobel Prize in Physics for his work on induced radioactivity by neutron bombardment and the discovery of transuranic elements. He was widely regarded as one of the very few physicists to excel both theoretically and experimentally.

Fermi's first major contribution was to statistical mechanics. After Wolfgang Pauli announced his exclusion principle in 1925, Fermi followed with a paper in which he applied the principle to an ideal gas, employing a statistical formulation now known as Fermi–Dirac statistics. Today, particles that obey the exclusion principle are called "fermions". Later Pauli postulated the existence of an uncharged invisible particle emitted along with an electron during beta decay, to satisfy the law of conservation of energy. Fermi took up this idea, developing a model that incorporated the postulated particle, which he named the "neutrino". His theory, later referred to as Fermi's interaction and still later as weak interaction, described one of the four fundamental forces of nature. Through experiments inducing radioactivity with recently discovered neutrons, Fermi discovered that slow neutrons were more easily captured than fast ones, and developed the Fermi age equation to describe this. After bombarding thorium and uranium with slow neutrons, he concluded that he had created new elements; although he was awarded the Nobel Prize for this discovery, the new elements were subsequently revealed to be fission products.

Fermi left Italy in 1938 to escape new Italian Racial Laws that affected his Jewish wife Laura. He emigrated to the United States where he worked on the Manhattan Project during World War II. Fermi led the team that designed and built Chicago Pile-1, that went critical on 2 December 1942, demonstrating the first artificial self-sustaining nuclear chain reaction. He was on hand when the X-10 Graphite Reactor at Oak Ridge, Tennessee went critical in 1943, and when the B Reactor at the Hanford Site did so the next year. At Los Alamos he headed F Division, part of which worked on Edward Teller's thermonuclear "Super" bomb. He was present at the Trinity test on 16 July 1945, where he used his Fermi method to estimate the bomb's yield.

After the war, Fermi served under Oppenheimer on the influential General Advisory Committee, which advised the Atomic Energy Commission on nuclear matters and policy. Following the detonation of the first Soviet fission bomb in August 1949, he strongly opposed the development of a hydrogen bomb on both moral and technical grounds. He was among the scientists who testified on Oppenheimer's behalf at the 1954 hearing that resulted in the denial of the latter's security clearance. Fermi did important work in particle physics, especially related to pions and muons, and he speculated that cosmic rays arose through material being accelerated by magnetic fields in interstellar space. Many awards, concepts, and institutions are named after Fermi, including the Enrico Fermi Award, the Enrico Fermi Institute, the Fermi National Accelerator Laboratory, the Fermi Gamma-ray Space Telescope, the Enrico Fermi Nuclear Generating Station, and the synthetic element fermium.

Contents [hide]

Enrico Fermi

Enrico Fermi (1901–1954)

Born	29 September 1901 Rome, Italy
Died	28 November 1954 (aged 53) Chicago, United States
Citizenship	Italy (1901–54) United States (1944–54)
Fields	Physics
Institutions	Scuola Normale Superiore University of Göttingen Leiden University University of Florence

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Print/export

Languages
Alemannisch
العربية
Aragonés
Asturianu
Azərbaycanca
বাংলা
Bân-lâm-gú
Башкортса
Беларуская
Беларуская (тарашкевіца)
Български

Noam Chomsky

From Wikipedia, the free encyclopedia

"Chomsky" redirects here. For other uses, see Chomsky (disambiguation).

Avram Noam Chomsky (/ˈnoʊm ˈtʃɒmski/; born December 7, 1928) is an American linguist, philosopher,[20][21] cognitive scientist, logician,[22][23][24] political commentator and activist. Sometimes described as the "father of modern linguistics",[25][26] Chomsky is also a major figure in analytic philosophy.[20] He has spent most of his career at the Massachusetts Institute of Technology (MIT), where he is currently Professor Emeritus, and has authored over 100 books. He has been described as a prominent cultural figure, and was voted the "world's top public intellectual" in a 2005 poll.[27]

Born to a middle-class Ashkenazi Jewish family in Philadelphia, Chomsky developed an early interest in anarchism from relatives in New York City. He later undertook studies in linguistics at the University of Pennsylvania, where he obtained his BA, MA, and PhD, while from 1951 to 1955 he was appointed to Harvard University's Society of Fellows. In 1955 he began work at MIT, soon becoming a significant figure in the field of linguistics for his publications and lectures on the subject. He is credited as the creator or co-creator of the Chomsky hierarchy, the universal grammar theory, and the Chomsky–Schützenberger theorem. In 1967 he gained public attention for his vocal opposition to U.S. involvement in the Vietnam War, in part through his essay *The Responsibility of Intellectuals*, and came to be associated with the New Left while being arrested on multiple occasions for his anti-war activism. While expanding his work in linguistics over subsequent decades, he also developed the propaganda model of media criticism with Edward S. Herman. Following his retirement from active teaching, he has continued his vocal public activism, praising the Occupy movement for example.

Chomsky has been a highly influential academic figure throughout his career, and was cited within the field of Arts and Humanities more often than any other living scholar between 1980 and 1992. He was also the eighth most cited scholar overall within the Arts and Humanities Citation Index during the same period.[28][29][30][31] His work has influenced fields such as artificial intelligence, cognitive science, computer science, logic, mathematics, music theory and analysis, political science, programming language theory and psychology.[30][31][32][33][34] Chomsky continues to be well known as a political activist, and a leading critic of U.S. foreign policy, state capitalism, and the mainstream news media. Ideologically, he aligns himself with anarcho-syndicalism and libertarian socialism.

Contents [hide]

Noam Chomsky

Noam Chomsky in 2005

Other names	Avram Noam Chomsky
Born	December 7, 1928 (age 85) Philadelphia, Pennsylvania, United States
Era	20th / 21st-century philosophy
Region	Western philosophy
School	Generative linguistics, Analytic philosophy
Main interests	Linguistics · Metalinguistics

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Print/export

Languages
Afrikaans
Alemannisch
العربية
বাংলা
Bân-lâm-gú
Беларуская
Български
Català
Čeština
Dansk
Deutsch
Eesti

Francis Crick

From Wikipedia, the free encyclopedia

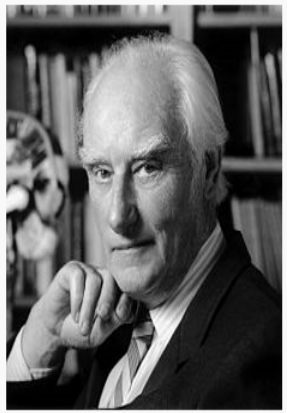
Francis Harry Compton Crick, OM, FRS (8 June 1916 – 28 July 2004) was an English molecular biologist, biophysicist, and neuroscientist, most noted for being a co-discoverer of the structure of the DNA molecule in 1953 with James Watson. He, Watson, and Maurice Wilkins were jointly awarded the 1962 Nobel Prize for Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".[1][2]

Crick was an important theoretical molecular biologist and played a crucial role in research related to revealing the genetic code. He is widely known for use of the term "central dogma" to summarize an idea that genetic information flow in cells is essentially one-way, from DNA to RNA to protein.[3]

During the remainder of his career, he held the post of J.W. Kieckhefer Distinguished Research Professor at the Salk Institute for Biological Studies in La Jolla, California. His later research centered on theoretical neurobiology and attempts to advance the scientific study of human consciousness. He remained in this post until his death; "he was editing a manuscript on his death bed, a scientist until the bitter end" according to Christof Koch.[4]

Contents [hide]

Francis Crick



Francis Crick

Born	Francis Harry Compton Crick 8 June 1916 Weston Favell, Northamptonshire, England, UK
Died	28 July 2004 (aged 88) San Diego, California, U.S. Colon cancer
Residence	UK, U.S.
Nationality	British
Fields	Physics Molecular biology
Institutions	University of Cambridge University College London Cavendish Laboratory MRC Laboratory of Molecular Biology Salk Institute for Biological

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Print/export

Languages
Alemannisch
العربية
বাংলা
Bân-lâm-gú
Беларуская
Български
Bosanski
Català
Čeština
Dansk
Deutsch
Eesti

James Watson

From Wikipedia, the free encyclopedia

For other people named James Watson, see James Watson (disambiguation).

James Dewey Watson, KBE (hon.), ForMemRS, (born April 6, 1928) is an American molecular biologist, geneticist and zoologist, best known as a co-discoverer of the structure of DNA in 1953 with Francis Crick. Watson, Crick, and Maurice Wilkins were awarded the 1962 Nobel Prize in Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".[4]

After studies at the University of Chicago (B.S., 1947) and Indiana University (Ph.D., 1950), did postdoctoral research to absorb chemistry with the biochemist Herman Kalckar in Copenhagen, Watson worked at the University of Cambridge's Cavendish Laboratory in England, where he first met his future collaborator and friend Francis Crick.

From 1956 to 1976, Watson was on the faculty of the Harvard University Biology Department, promoting research in molecular biology.[4]

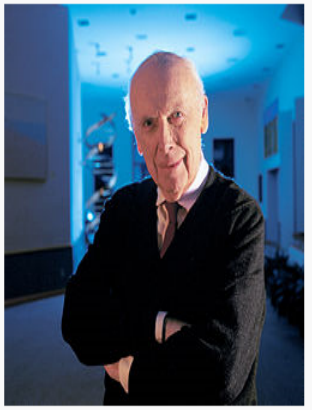
From 1968 Watson served as director of Cold Spring Harbor Laboratory (CSHL) on Long Island, New York, greatly expanding its level of funding and research. At CSHL, he shifted his research emphasis to the study of cancer, along with making it a world leading research center in molecular biology. In 1994, he started as president and served for 10 years. He was then appointed chancellor, serving until 2007.[5]

Between 1988 and 1992, Watson was associated with the National Institutes of Health, helping to establish the Human Genome Project.

Watson has written many science books, including the textbook *Molecular Biology of the Gene*[6] (1965) and his bestselling book *The Double Helix*[7] (1968) about the DNA structure discovery, reissued in a new edition in 2012 - *The Annotated and Illustrated Double Helix* edited by Alex Gann and Jan Witkowski.[8]

Contents [hide]

James Watson



James Watson

Born	James Dewey Watson April 6, 1928 (age 85) Chicago, Illinois, U.S.
Nationality	American
Fields	Genetics
Institutions	Indiana University Cold Spring Harbor Laboratory Harvard University University of Cambridge National Institutes of Health
Alma mater	University of Chicago Indiana University
Thesis	*The Biological Properties of X-Ray Inactivated Bacteriophage* (1951)

The Bacon Numbers

- Sometime around 1994, 3 students at Albright College in Pennsylvania adapted the idea of Erdos numbers to the collaboration graph of movie actors and actresses
  - Nodes are performers
  - An edge connects two performers if they've appeared together in a movie, and a performer's **Bacon number** is his or her distance in this graph to Kevin Bacon.
  - Using cast lists from the Internet Movie Database (IMDB), it is possible to compute Bacon numbers for all performers via breadth-first search, and as with mathematics, it's a small world indeed. (See next slide for the IMDbPY package)
  - The average Bacon number is approximately **3.002**, and it's a challenge to find one that's larger than 6.

IMDbPY is a Python package useful to retrieve and manage the data of the IMDb movie database about movies, people, characters and companies.

Download IMDbPY 5.1

Documentation

- Written in pure Python (and few C lines).
- Platform-independent.
- Can retrieve data from both the IMDb's web server and a local copy of the whole database.
- Released under the terms of the GPL 2 license.
- A simple API.

#DOCUMENTATION

These are some of the documents for the current stable (5.1) IMDbPY release:

- A short introduction to IMDbPY, its features and instructions to install it (with notes for packagers)
- Instructions and tips to use IMDbPY on mobile systems
- How to put the whole IMDb's database in a SQL database
- Use export information in XML format
- Localize the tags used in the XML format
- Search and manage keywords
- Use the IMDbPY package in your programs

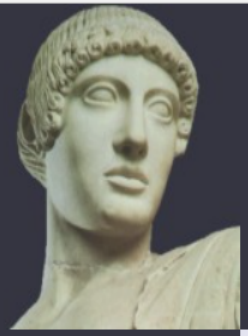
IMDBPy: An Example

```
from imdb import IMDb
ia = IMDb()

for t in ia.search_movie('la la land'):
    print t, t.movieID

lalaland = ia.get_movie('3783958')

print 'Cast:'
for i in range(len(lalaland['cast'])):
    print lalaland['cast'][i]['name'], 'as', lalaland['cast'][i].currentRole

print 'Director(s):'
for i in range(len(lalaland[‘director'])):
    print lalaland['director'][i]['name']

print 'Writer(s):'
for i in range(len(lalaland['writer'])):
    print lalaland['writer'][i]['name']
```

The Bacon Numbers

The Oracle of Bacon

oracleofbacon.org/movielinks.php



THE ORACLE OF BACON

Welcome
Credits
How it Works
Contact Us
Other stuff »

scarlett johansson has a Bacon number of 2. Find a different link

Scarlett Johansson
was in
He's Just Not That Into You (2009)
with
Jason Moffett
was in
My One and Only (2009)
with
Kevin Bacon

Kevin Bacon to scarlett johanss Find link More options >>

A Recent Facebook Experiment

- A recent study by researchers from Facebook and the University of Milan looked at 721 million Facebook users, who had 69 billion unique friendships among them, and revealed an average distance of **4.74** (3.74 intermediaries).

*(Backstrom, L., Boldi, P., Rosa, M., Ugander, J., and Vigna, S., "**Four degrees of separation**," http://arxiv.org/abs/1111.4570, Jan. 6, 2012.)*

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.

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

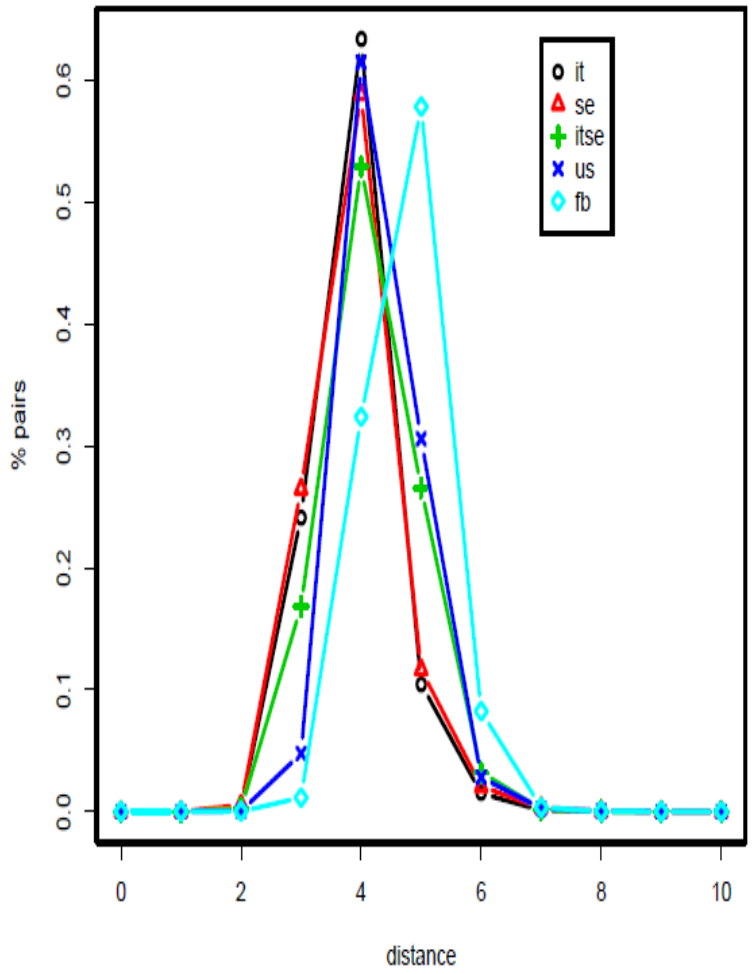


Figure 2: The probability mass functions of the distance distributions of the current graphs (truncated at distance 10).

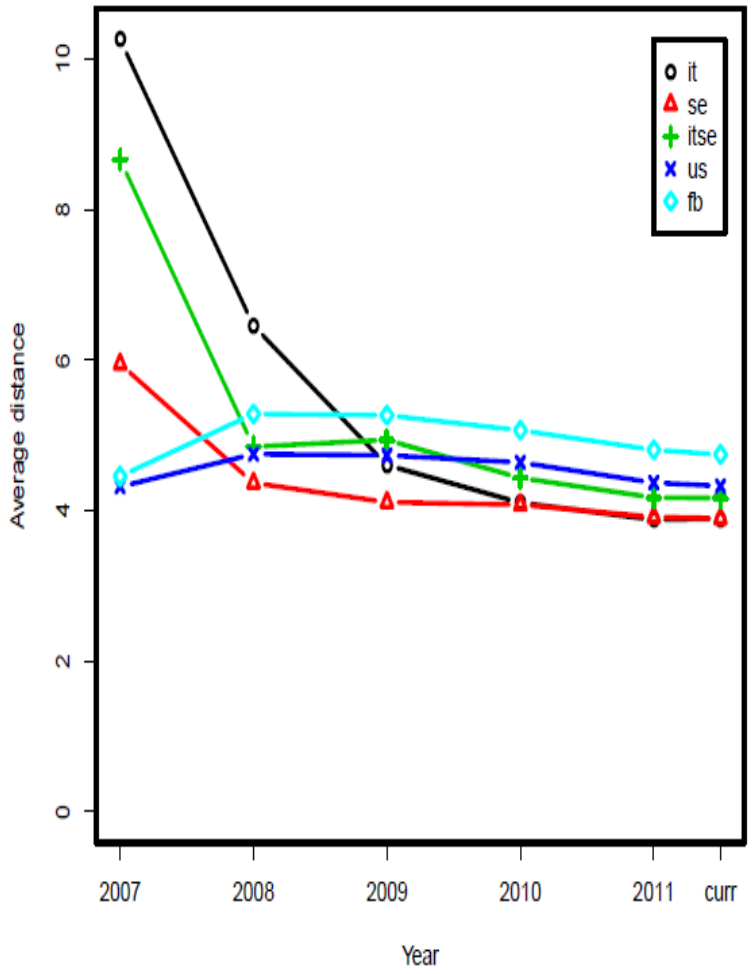


Figure 3: The average distance graph. See also Table 6.

A Recent Twitter Experiment

- Another recent study finds an average degree of separation of 3.43 between two random Twitter users.

*(Bakhshandeh, R., Samadi, M., Azimifar, Z., and Schaeffer, J., "**Degrees of separation in social networks**," Proceedings of the Fourth International Symposium on Combinatorial Search, Barcelona, Spain, July 15–16, 2011.)*

Proceedings, The Fourth International Symposium on Combinatorial Search (SoCS-2011)

Degrees of Separation in Social Networks

Reza Bakhshandeh
Shiraz University
Shiraz, Iran
bakhshandeh@cse.shirazu.ac.ir

Mehdi Samadi
Carnegie Mellon University
Pittsburgh, United States
msamadi@cs.cmu.edu

Zohreh Azimifar
Shiraz University
Shiraz, Iran
azimifar@cse.shirazu.ac.ir

Jonathan Schaeffer
University of Alberta
Edmonton, Canada
jonathan@cs.ualberta.ca

Abstract

Social networks play an increasingly important role in today's society. Special characteristics of these networks make them challenging domains for the search community. In particular, social networks of users can be viewed as search graphs of nodes, where the cost of obtaining information about a node can be very high. This paper addresses the search problem of identifying the degree of separation between two users. New search techniques are introduced to provide optimal or near-optimal solutions. The experiments are performed using Twitter, and they show an improvement of several orders of mag-

Each node corresponds to a user in the social network, with an edge showing direct connectivity of two users. Thus the degree of separation between two users becomes a search for the minimum cost connection between them. Computing the degree of separation in a social network is an unusual search problem. The tree has a shallow search depth but can have an enormous branching factor (the number of "friends" that each user has). The cost of obtaining the information on the search graph means issuing a query over the Internet (which is many orders of magnitude slower than CPU-based search). An unwelcome complication is that most social net-

Row	Heuristic for tie-breaking	Hits (out of 1,500)	Degree of Separation	σ	Node Generation (average)	Twitter Requests (average)	Time (minutes) (average)
Bidirectional Search: Breadth-first (optimal)							
1	Max #followings	1480	3.43	0.68	337	371	8.1 (128)
Bidirectional Search: Probabilistic							
2	Max #followings	1,500	3.880	0.77	204	13.3	0.3 (0.3)
Bounded Bidirectional Search (optimal)							
3	Max #followings	1,500	3.435	0.67	222	67	1.4 (1.4)

Table 2: Experimental results on 1,500 random pairs chosen from Twitter.

to reduce the cost of obtaining an optimal solution. First, run the probabilistic search, obtaining a solution of length K. Second, run the breadth-first bidirectional search to a maximum combined depth of $K - 1$. If this latter search is successful then the optimal solution has been found. If it fails, then the probabilistic search result of K is optimal.

This technique, called Bounded Bidirectional Search, can significantly reduce the number of Twitter requests needed

References

Adamic, L. A., and Adar, E. 2005. How to search a social network. *Social Networks* 27:2005.

Adamic, L. A.; Lukose, R. M.; Puniyani, A. R.; and Huberman, B. A. 2001. Search in power-law networks. *PHYS.REV.E* 64:046135.

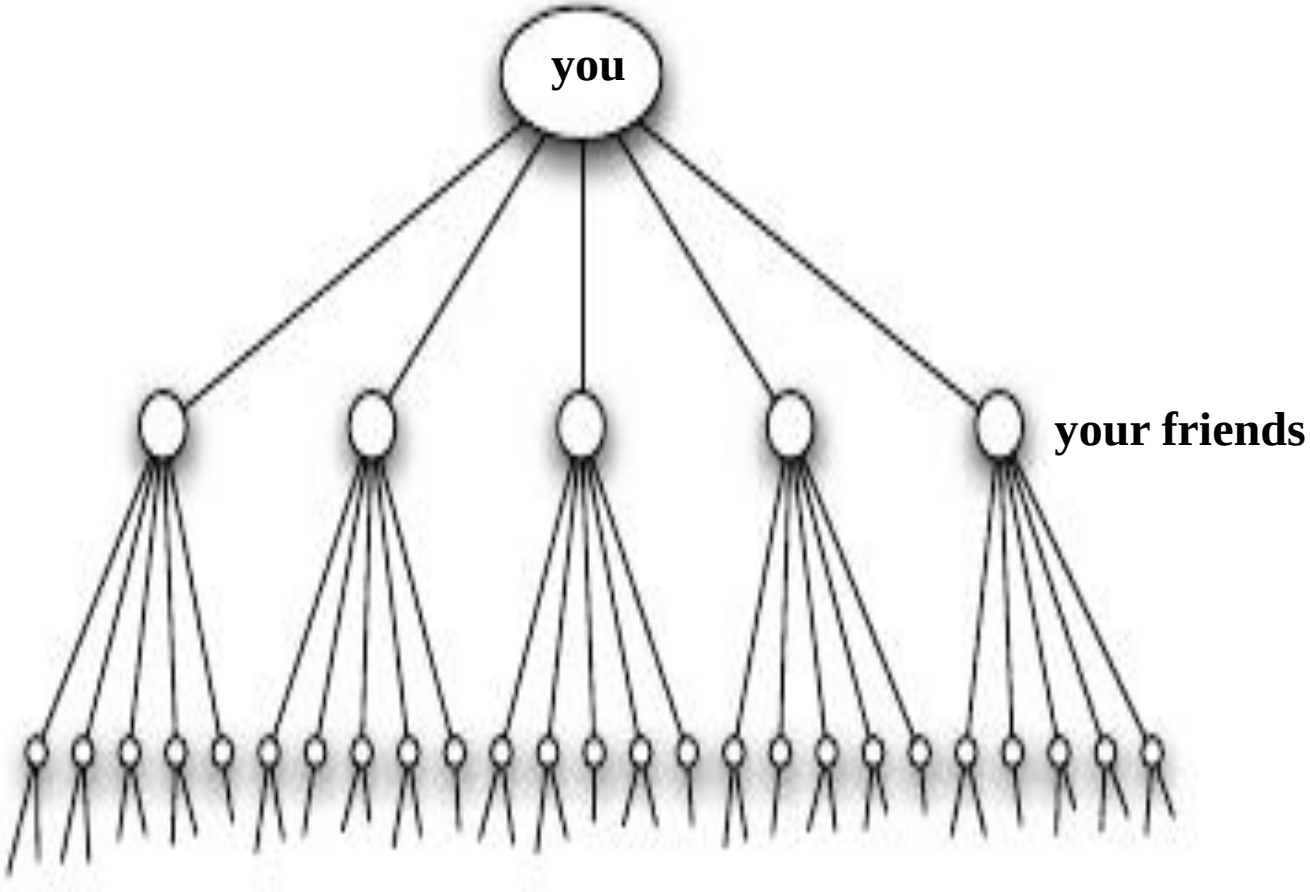
Adamic, L. A.; Lukose, R. M.; and Huberman, B. A. 2003.

Why Is the World Small?

- Milgram's experiment demonstrated two striking facts about large social networks:
  - Short paths are there (in abundance)
  - People, acting without any sort of global "map" of the network, are effective at collectively finding these short paths.
- It is easy to imagine a social network where the first is true but the second isn't
  - A large social-networking site where everyone was known only by 9-digit id's would be such a social network:
    - If you were told, "Forward this letter to user **482285204**, using only people you know on a first-name basis," the task would clearly be hopeless.

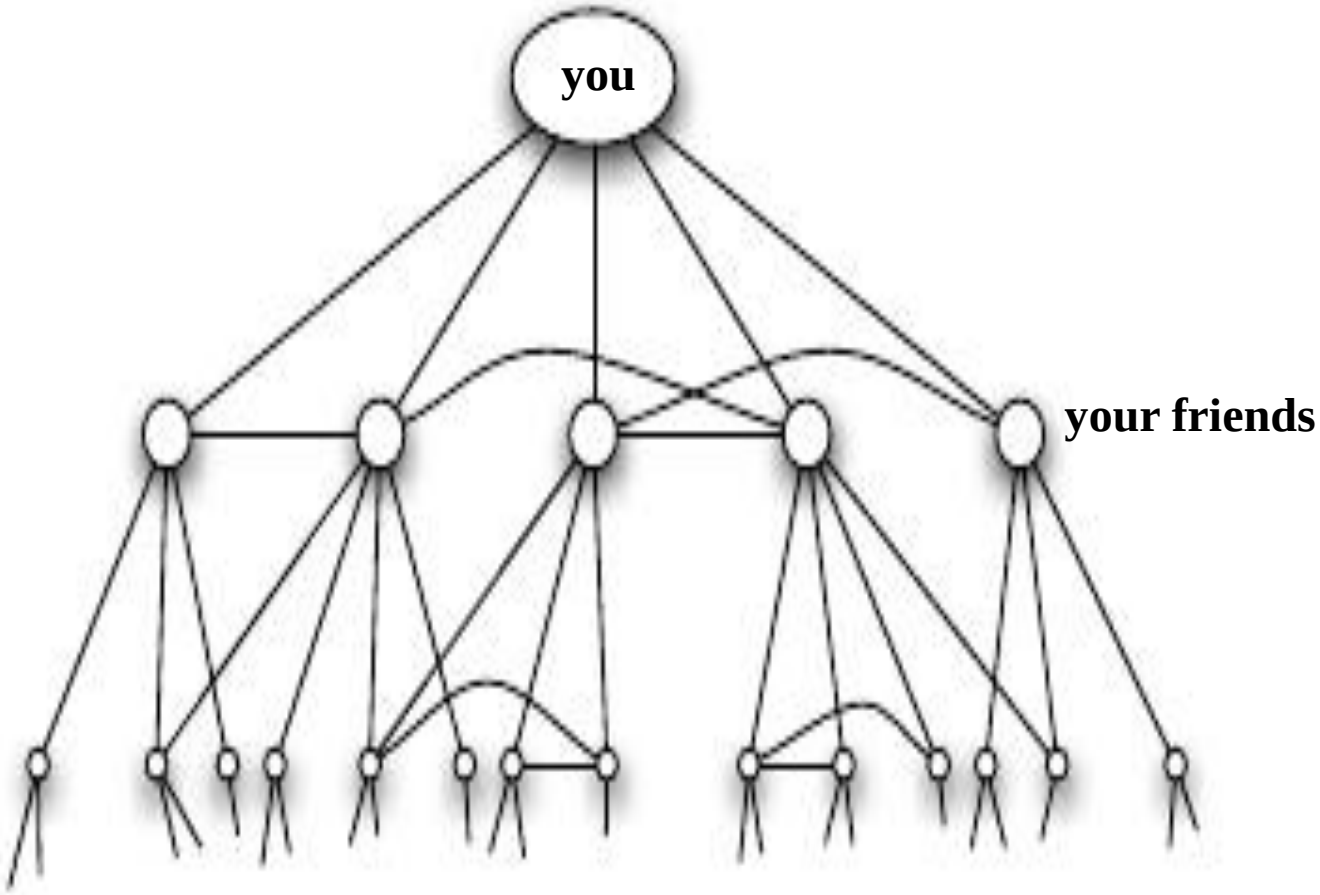
Models for Short Paths

(a) Pure **exponential growth** produces a small world



Models for Short Paths

(b) **Triadic closure** reduces the growth rate



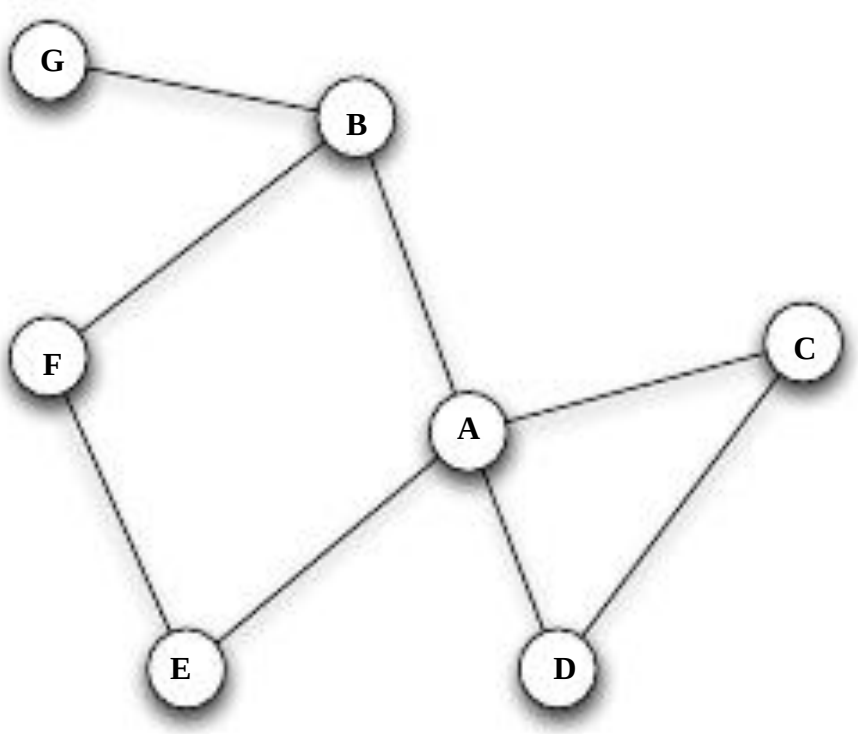
The Triadic Closure Principle

- So far we have treated networks as static structures, but in reality they evolve over time.
  - People (or nodes) come and go…
  - Relationships (or edges) too…
  - What are the mechanisms that make them come and go?
  - The precise answer will of course vary depending on the type of network, but one of the most basic principles is the following:

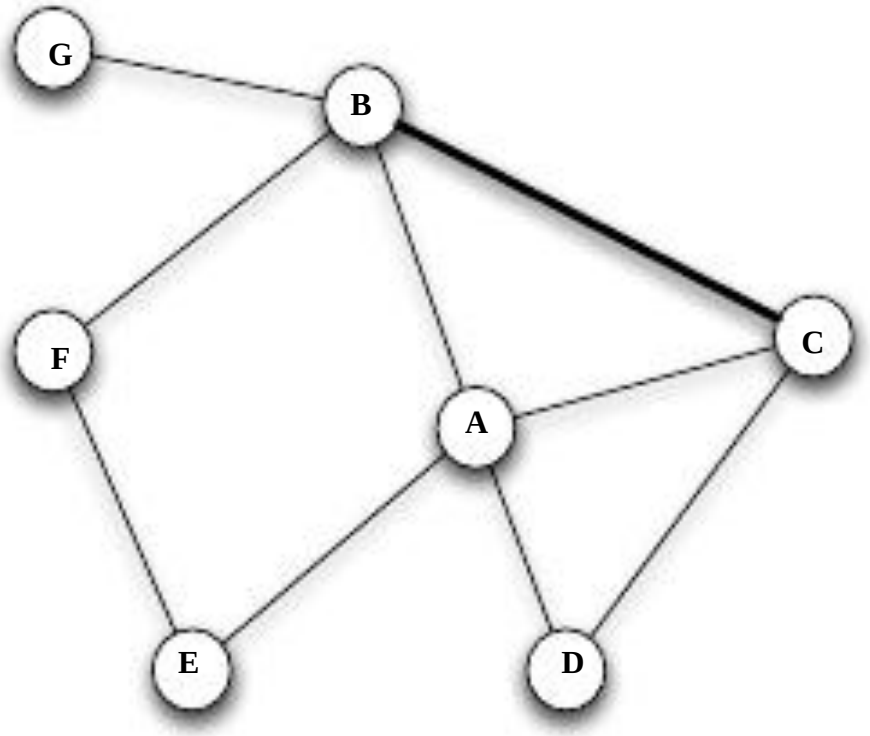
    > If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.

- This principle is called the triadic closure principle

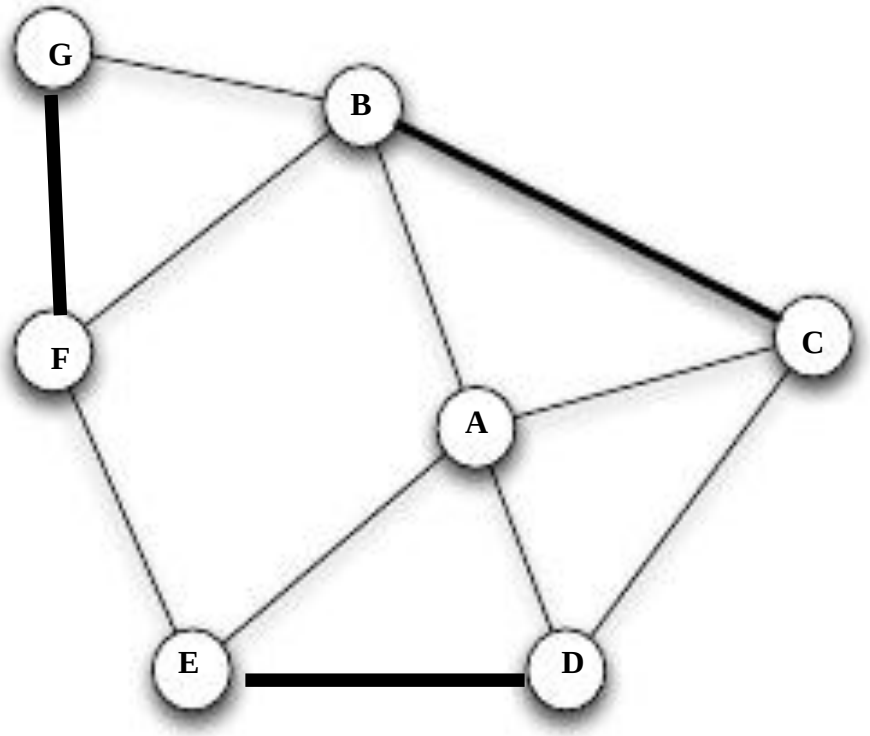
The Triadic Closure Principle



The Triadic Closure Principle



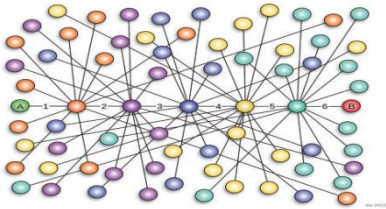
The Triadic Closure Principle



The Clustering Coefficient

- The basic role of triadic closure in social networks has motivated the formulation of simple social network measures to capture its prevalence.
  - One of them is **clustering coefficient**.
- The clustering coefficient of a node A is defined as:
  - the probability that two randomly selected friends of A are friends with each other.
  - Or, in more practical terms, the fraction of pairs of A's friends that are connected to each other by edges.
- In general, the clustering coefficient of a node ranges from 0 (when none of the node's friends are friends with each other) to 1 (when all of the node's friends are friends with each other).

Network Modeling (Revisited)

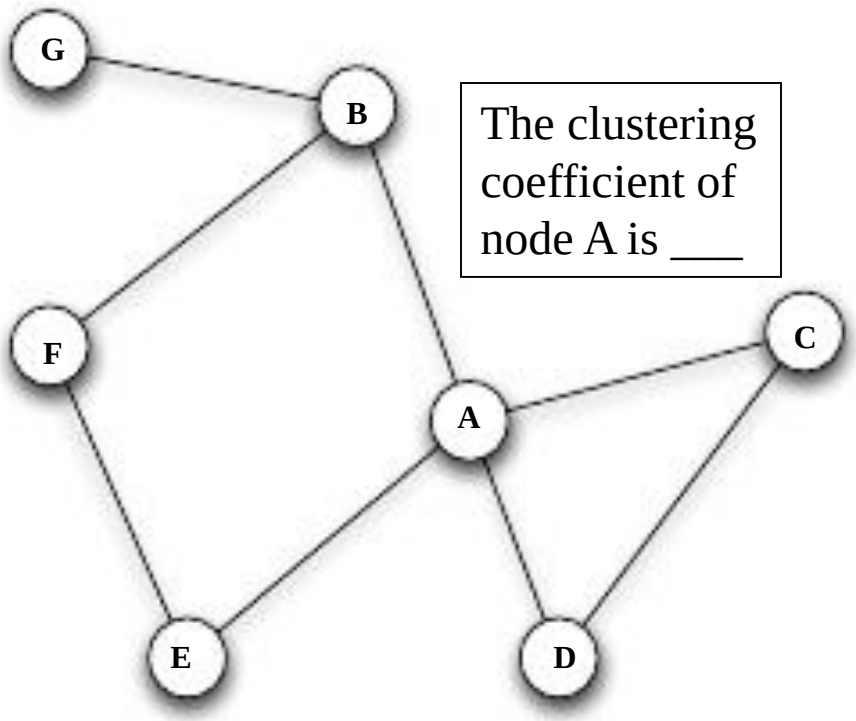


- Large Networks demonstrate statistical patterns:
  - Small-world phenomenon (6 degrees of separation)
  - Power-law distribution (a.k.a. scale-free distribution)

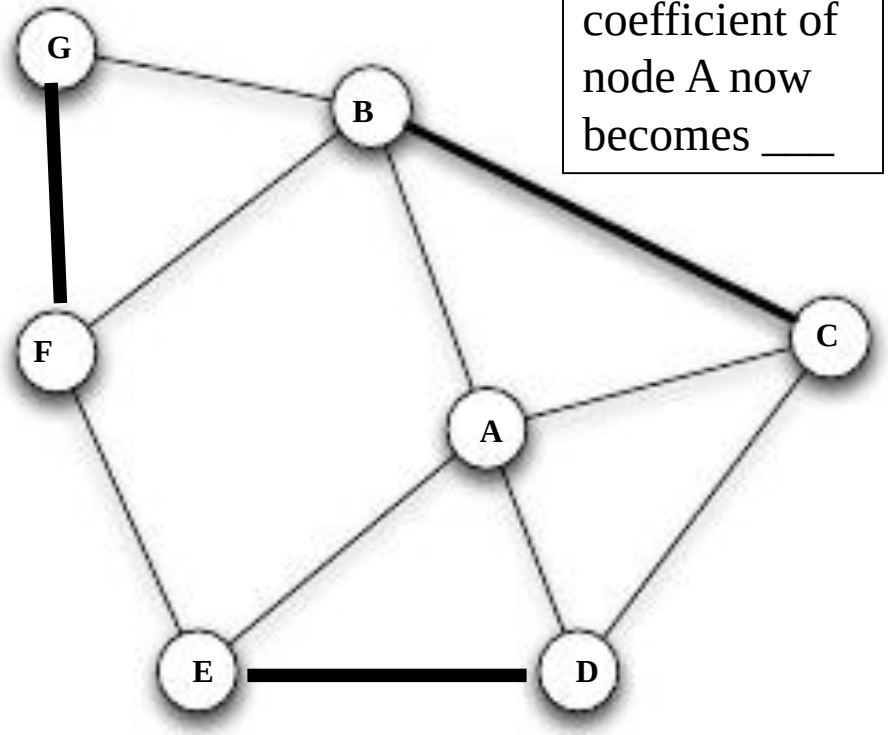
$$f(x) = ax^k \qquad f(cx) = a(cx)^k = c^k f(x) \propto f(x).$$

  - Community structure (**high clustering coefficient**)
- Model the network dynamics
  - Find a mechanism such that the statistical patterns observed in large-scale networks can be reproduced.
  - Examples: random graph, preferential attachment process, Watts and Strogatz model
- Used for simulation to understand network properties
  - Thomas Shelling's famous simulation: What could cause the segregation
  - Network robustness under attack
  - Information diffusion within a given network structure

The Triadic Closure Principle

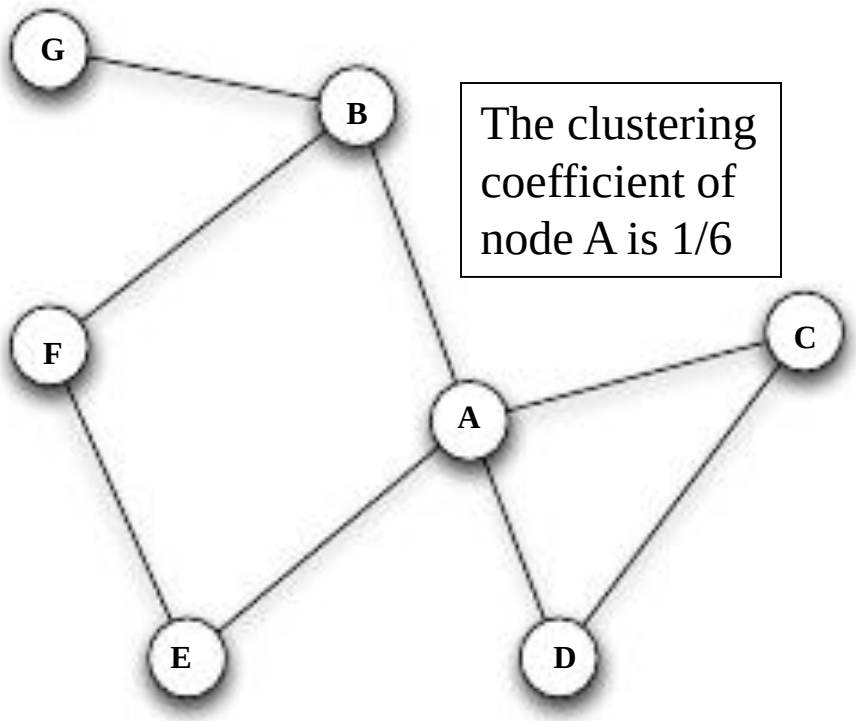


The clustering coefficient of node A is ___

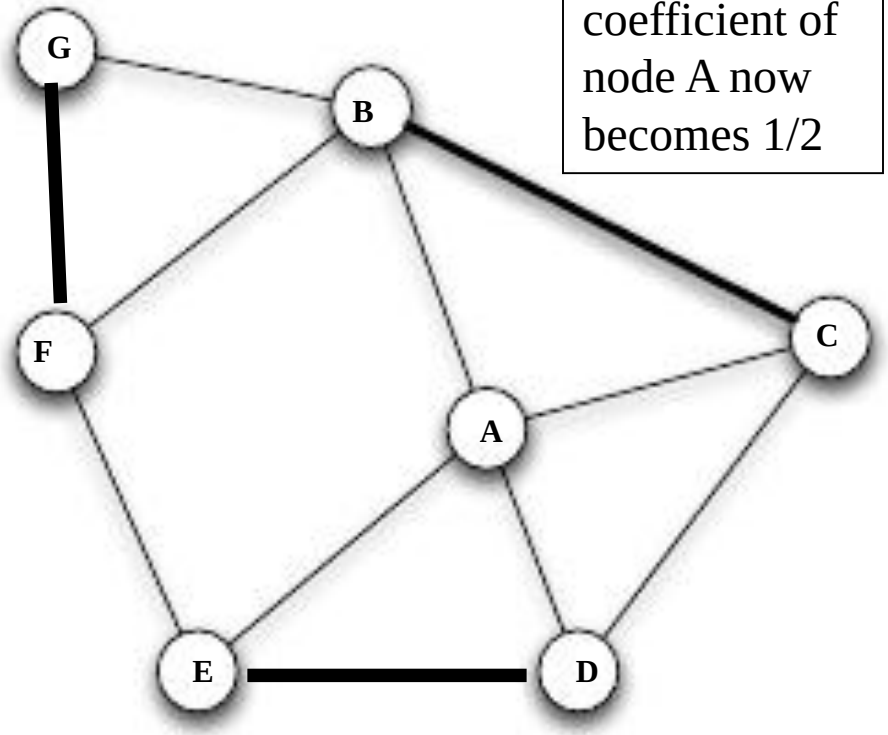


The clustering coefficient of node A now becomes ___

The Triadic Closure Principle



The clustering coefficient of node A is 1/6



The clustering coefficient of node A now becomes 1/2

By *Jure Leskovec*

STANFORD UNIVERSITY

SNAP for C++
SNAP for Python
SNAP Datasets
What's new
People
Papers
Projects
Citing SNAP
Links
About
Contact us

Open positions

We have filled all the positions for this quarter. More info.

Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks, articles, and metadata : Talk, editing, voting, and article data from Wikipedia
- Temporal networks : networks where edges have timestamps
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also availalbe from UF Sparse Matrix collection. **Visualizations of SNAP networks** by Tim Davis.

Social networks

Name	Type	Nodes	Edges	Description
ego-Facebook	Undirected	4,039	88,234	Social circles from Facebook (anonymized)
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
ego-Twitter	Directed	81,306	1,768,149	Social circles from Twitter

By *Jure Leskovec*

STANFORD UNIVERSITY

Stanford Network Analysis Project

Open positions

Open research positions in **SNAP** group are available here.

SNAP for C++: Stanford Network Analysis Platform

Stanford **N**etwork **A**nalysis **P**latform (**SNAP**) is a general purpose network analysis and graph mining library. It is written in C++ and easily scales to massive networks with hundreds of millions of nodes, and billions of edges. It efficiently manipulates large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges. SNAP is also available through the NodeXL which is a graphical front-end that integrates network analysis into Microsoft Office and Excel.

Snap.py: SNAP for Python

Snap.py is a Python interface for SNAP. It provides performance benefits of SNAP, combined with flexibility of Python. Most of the SNAP C++ functionality is available via Snap.py in Python.

Stanford Large Network Dataset Collection

A collection of more than 50 large network datasets from tens of thousands of nodes and edges to tens of millions of nodes and edges. In includes social networks, web graphs, road networks, internet networks, citation networks, collaboration networks, and communication networks.

Tutorials

By *Jure Leskovec*

STANFORD UNIVERSITY

Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks, articles, and metadata : Talk, editing, voting, and article data from Wikipedia
- Temporal networks : networks where edges have timestamps
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also availalbe from UF Sparse Matrix collection. **Visualizations of SNAP networks** by Tim Davis.

SNAP for C++
SNAP for Python
SNAP Datasets
What's new
People
Papers
Projects
Citing SNAP
Links
About
Contact us

Open positions

We have filled all the positions for this quarter. More info.

Social networks

Name	Type	Nodes	Edges	Description
ego-Facebook	Undirected	4,039	88,234	Social circles from Facebook (anonymized)
ego-Gplus	Directed	107,614	13,673,453	Social circles from Google+
ego-Twitter	Directed	81,306	1,768,149	Social circles from Twitter

SNAP for C++
SNAP for Python
SNAP Datasets
What's new
People
Papers
Projects
Citing SNAP
Links
About
Contact us

Open positions

We have filled all the positions for this quarter. More info.

Dataset information

This dataset consists of 'circles' (or 'friends lists') from Facebook. Facebook data was collected from survey participants using this Facebook app. The dataset includes node features (profiles), circles, and ego networks.

Facebook data has been anonymized by replacing the Facebook-internal ids for each user with a new value. Also, while feature vectors from this dataset have been provided, the interpretation of those features has been obscured. For instance, where the original dataset may have contained a feature "political=Democratic Party", the new data would simply contain "political=anonymized feature 1". Thus, using the anonymized data it is possible to determine whether two users have the same political affiliations, but not what their individual political affiliations represent.

Data is also available from Google+ and Twitter.

Dataset statistics	
Nodes	4039
Edges	88234
Nodes in largest WCC	4039 (1.000)
Edges in largest WCC	88234 (1.000)
Nodes in largest SCC	4039 (1.000)
Edges in largest SCC	88234 (1.000)
Average clustering coefficient	0.6055
Number of triangles	1612010
Fraction of closed triangles	0.2647
Diameter (longest shortest path)	8
90-percentile effective diameter	4.7

Note that these statistics were compiled by combining the ego-networks, including the ego nodes themselves (along with an edge to each of their friends).

Source (citation)

The Strong Triadic Closure Property

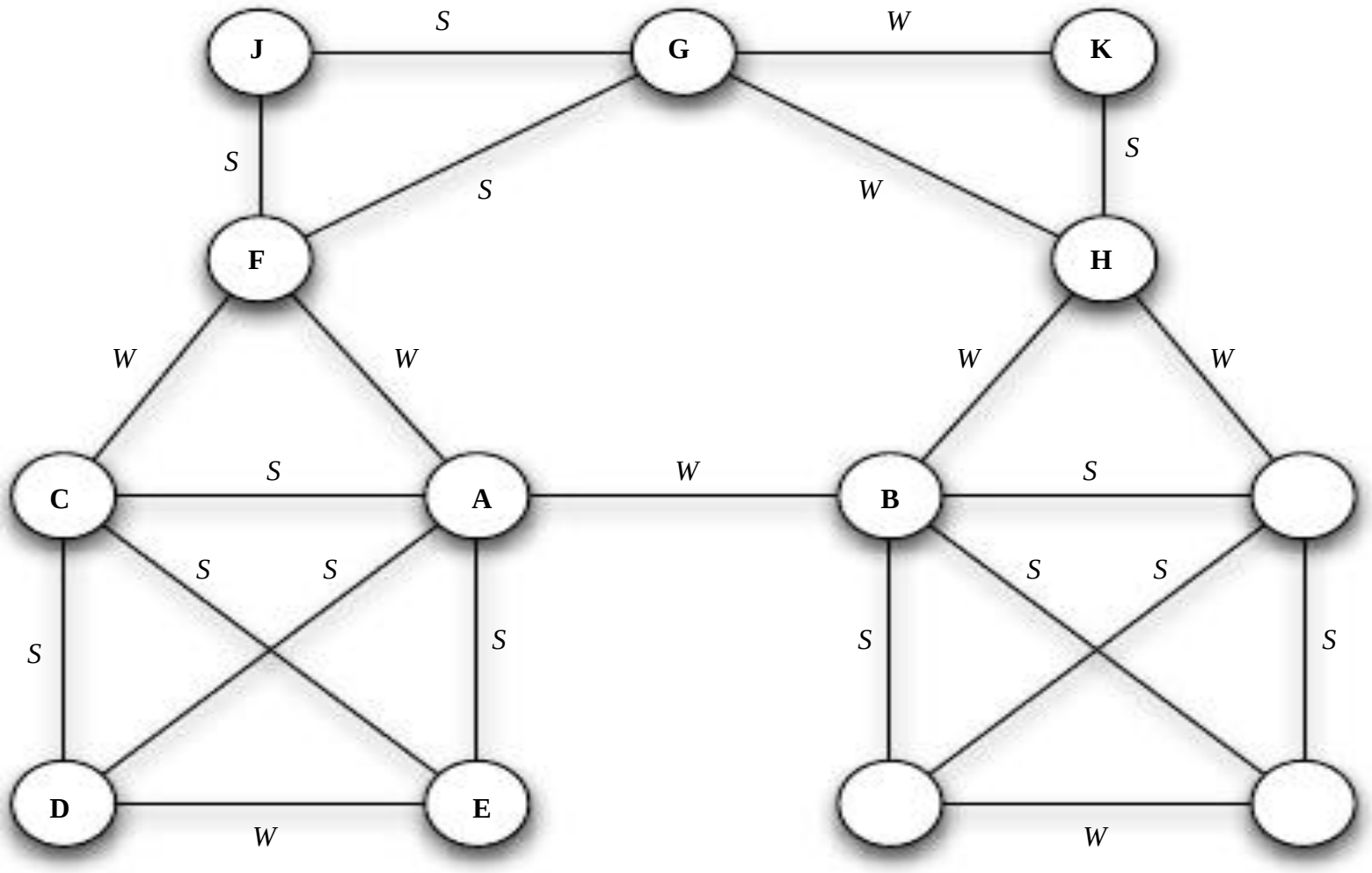
Observation:

If a node A has edges to nodes B and C, then the B-C edge is especially likely to form if A's edges to B and C are both **strong ties**.

Strong Ties & Weak Ties

- From our own experiences, we know that friendships (links in a social network) have different levels of strength.
  - Stronger links represent closer friendship and greater frequency of interaction.
  - For conceptual simplicity, and to match the friend/acquaintance dichotomy we usually use, we'll categorize all links in a social network as belonging to one of two types:
    - strong ties (the stronger links, corresponding to friends)
    - weak ties (the weaker links, corresponding to acquaintances)

Strong Ties & Weak Ties

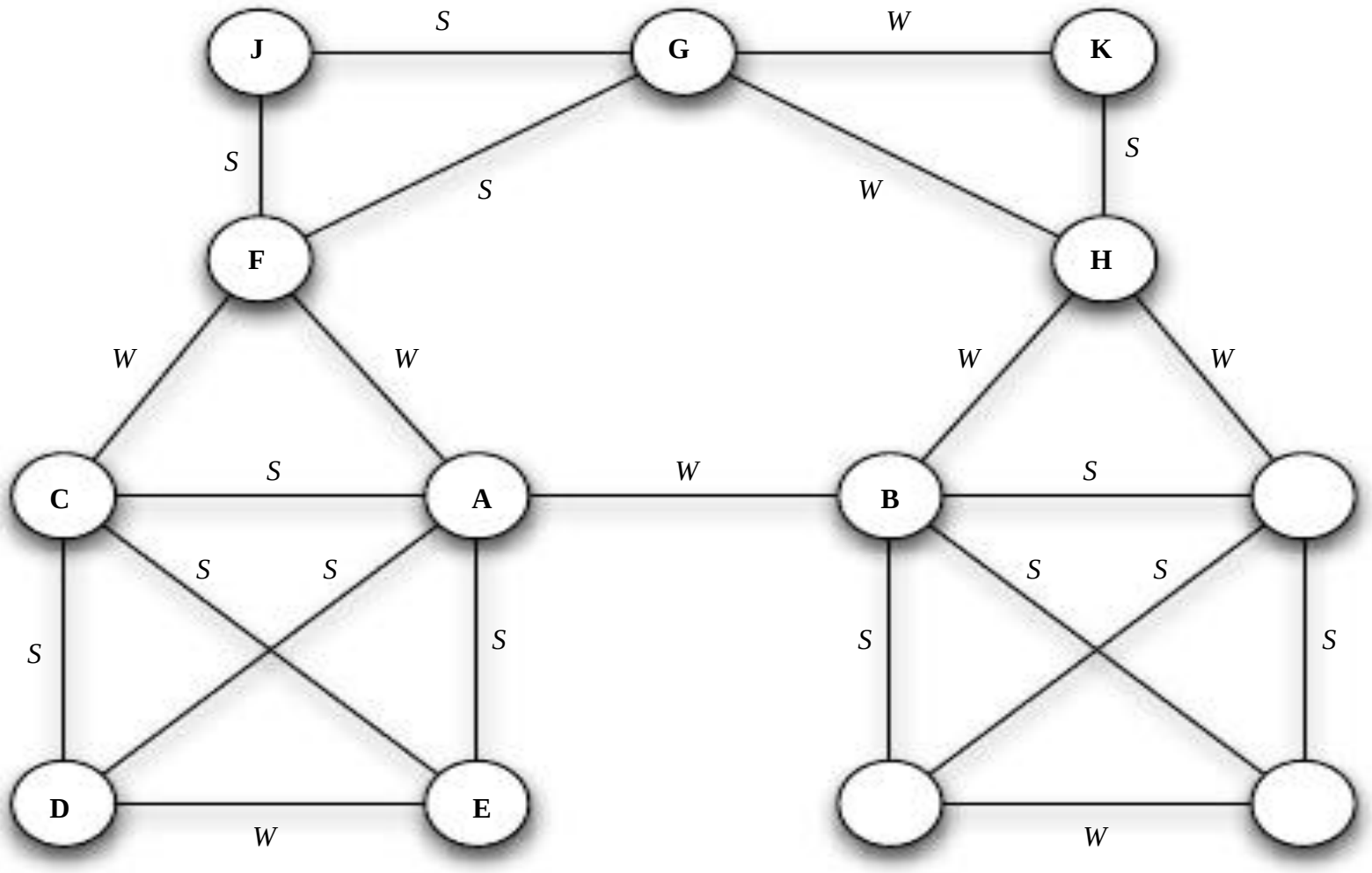


The Strong Triadic Closure Property

More formally:

- A node A violates the **Strong Triadic Closure Property** if it has strong ties to two other nodes B and C, and there is no edge at all (either a strong or weak tie) between B and C.
- A node A satisfies the **Strong Triadic Closure Property** if it does not violate it.

The Strong Triadic Closure Property



Does any node violate the strong triadic closure property?

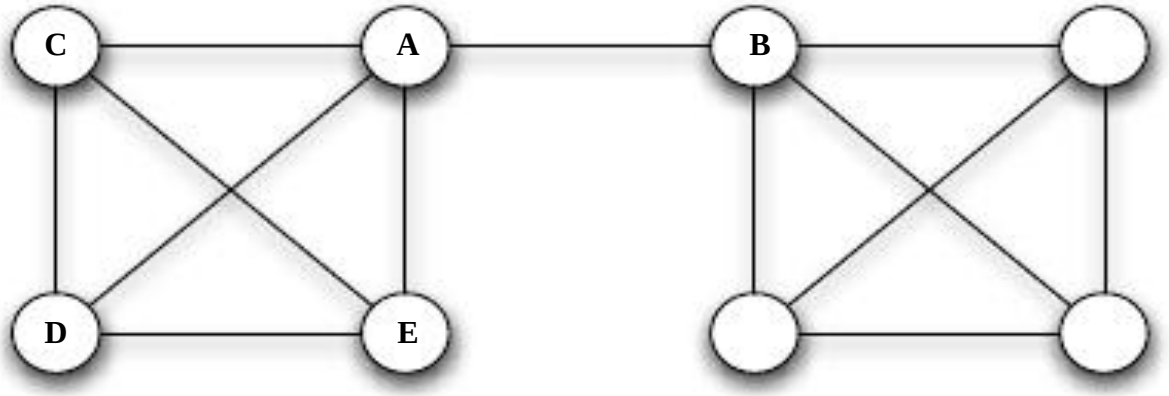
The Strong Triadic Closure Property

- The strong triadic closure property establishes a connection between:
  - The notion of **weak/strong ties** - a purely local, interpersonal concept, and
  - The notion of **local bridges -** a global, structural concept,
- by the following claim:

Claim: If a node A in a network satisfies the Strong Triadic Closure Property and is involved in at least two strong ties, then any local bridge it is involved in must be a weak tie.

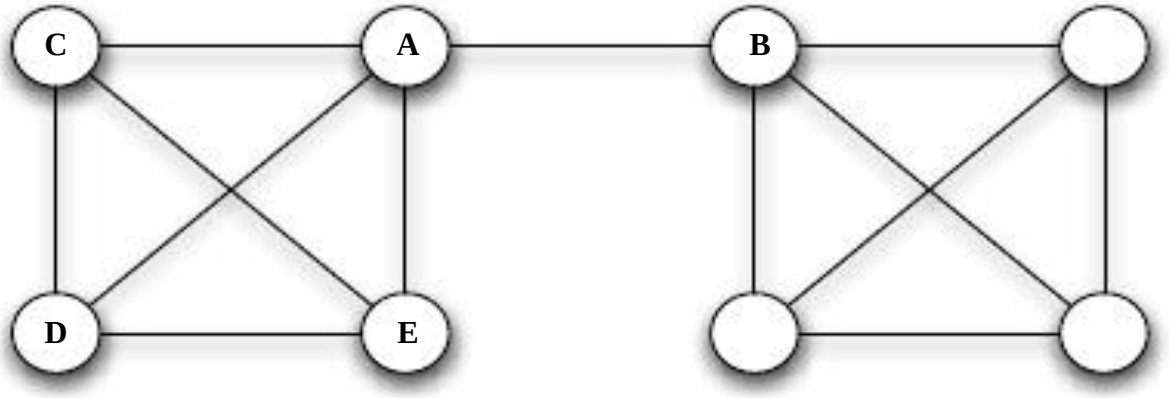
(See p.55, **Network, Crowds and Markets**, Chapter 3)

Bridges



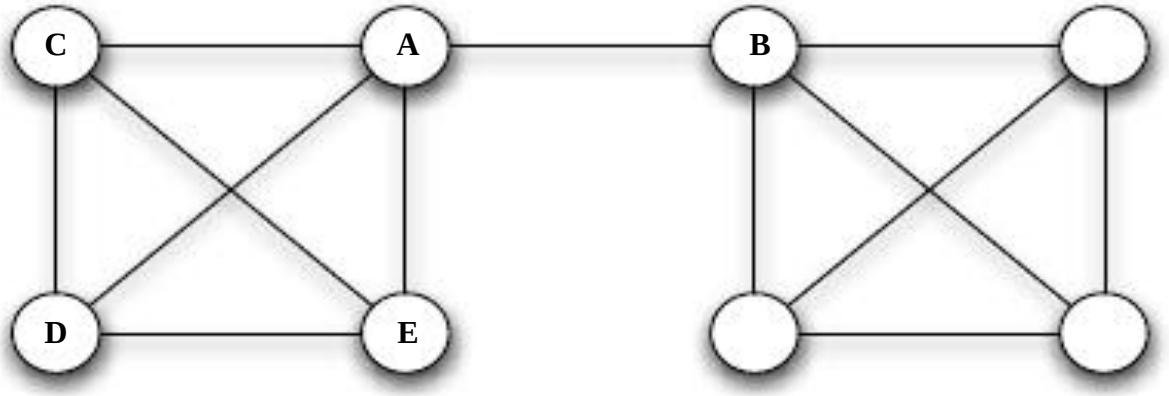
- A has four friends in this picture, but one of her friendships is qualitatively different from the others. Which one?

Bridges



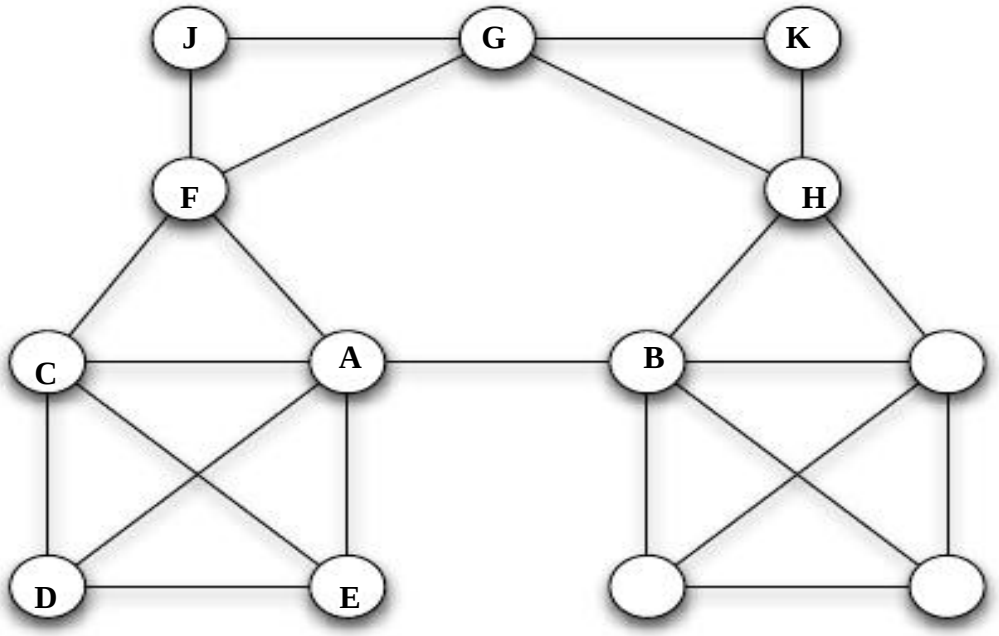
- A's links to C, D, and E connect her to a tightly-knit group of friends, while the link to B seems to reach into a different part of the network.
  - We could speculate that the structural peculiarity of the link to B will translate into differences in the role it plays in A's everyday life.

Bridges



- An edge joining two nodes A and B in a graph is a **bridge** if deleting the edge would cause A and B to lie in two different components.

Local Bridges



- An edge joining two nodes A and B in a graph is a **local bridge** if its endpoints A and B have no common friends.
  - In other words, if deleting the edge would increase the distance between A and B by at least two.
  - We say that the **span** of a local bridge is the distance its endpoints would be from each other if the edge were deleted.

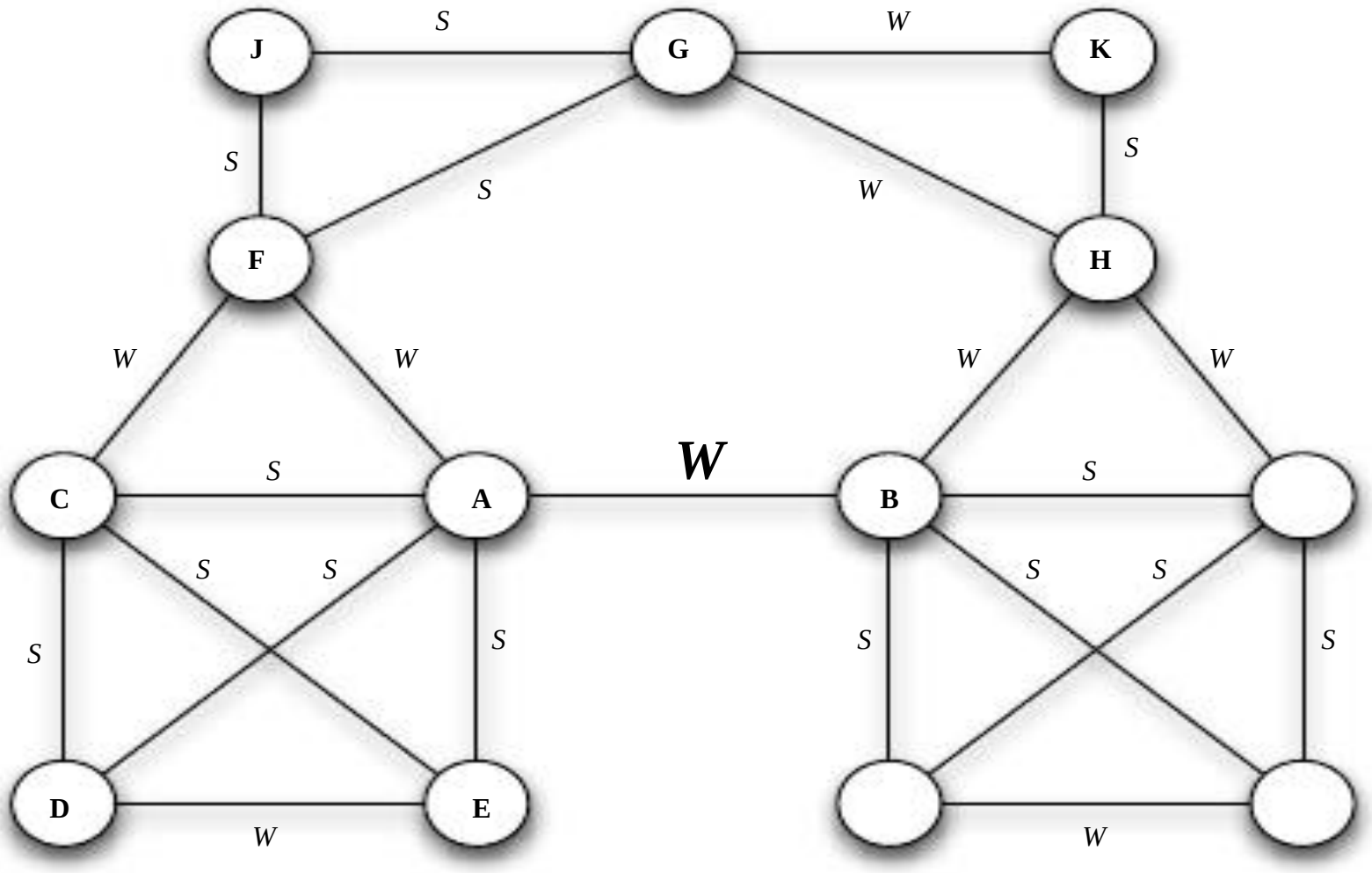
The Strong Triadic Closure Property

- The strong triadic closure property establishes a connection between:
  - The notion of **weak/strong ties** - a purely local, interpersonal concept, and
  - The notion of **local bridges -** a global, structural concept,
- by the following claim:

Claim: If a node A in a network satisfies the Strong Triadic Closure Property and is involved in at least two strong ties, then **any local bridge it is involved in must be a weak tie.**

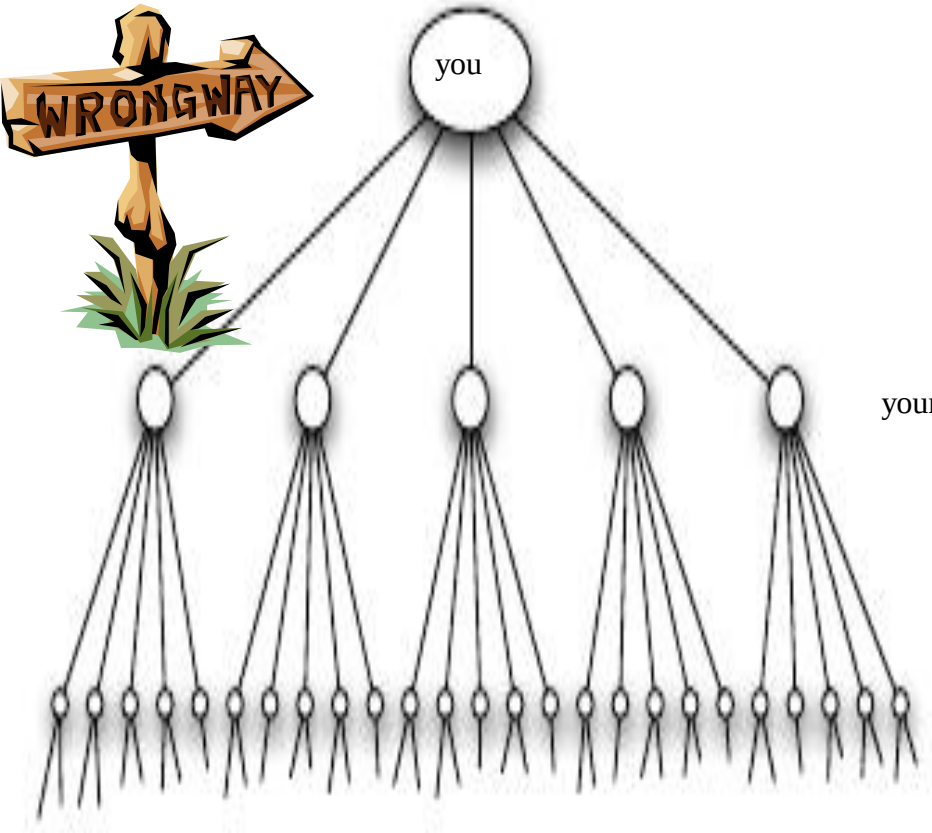
(See p.55, **Network, Crowds and Markets**, Chapter 3)

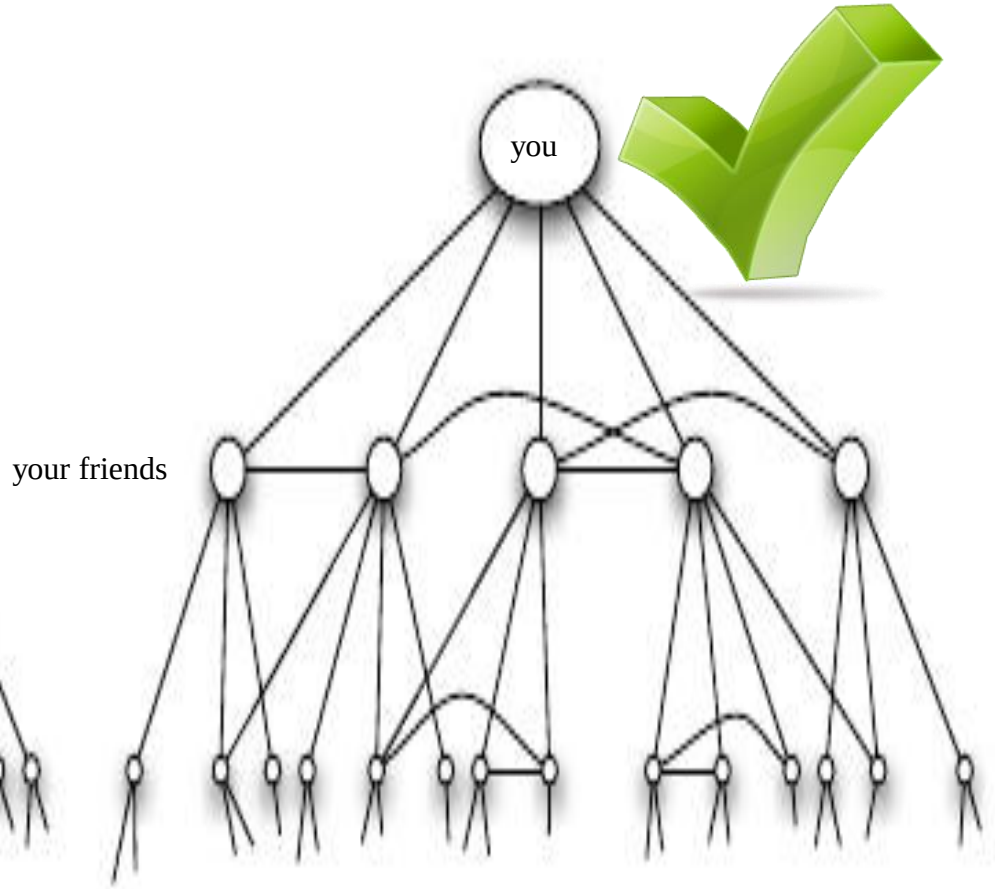
The Strong Triadic Closure Property



Models for Short Paths (Revisited)

(a) Pure **exponential growth** produces a small world





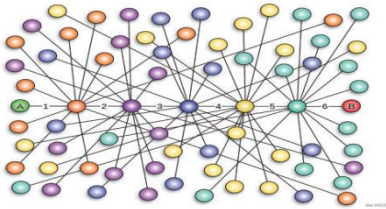
(b) **Triadic closure** reduces the growth rate

The Watts-Strogatz Model

- Can we make up a simple model that exhibits both of the features we've been discussing: many closed triads, but also very short paths?

- In 1998, Duncan Watts and Steve Strogatz argued that such a model follows naturally from a combination of two basic social-network ideas:
  - **Weak ties**: the links to acquaintances that connect us to parts of the network that would otherwise be far away
    - Weak ties produce very short paths
  - **Homophily**: the principle that we connect to others who are like ourselves
    - Homophily creates many closed triads

Network Modeling (Revisited)



- Large Networks demonstrate statistical patterns:
  - **Small-world phenomenon (6 degrees of separation)**
  - Power-law distribution (a.k.a. scale-free distribution)

$$f(x) = ax^k \qquad f(cx) = a(cx)^k = c^k f(x) \propto f(x).$$

  - Community structure (high clustering coefficient)
- Model the network dynamics
  - Find a mechanism such that the statistical patterns observed in large-scale networks can be reproduced.
  - Examples: random graph, preferential attachment process, **Watts and Strogatz model**
- Used for simulation to understand network properties
  - Thomas Shelling's famous simulation: What could cause the segregation
  - Network robustness under attack
  - Information diffusion within a given network structure

The Watts-Strogatz Model

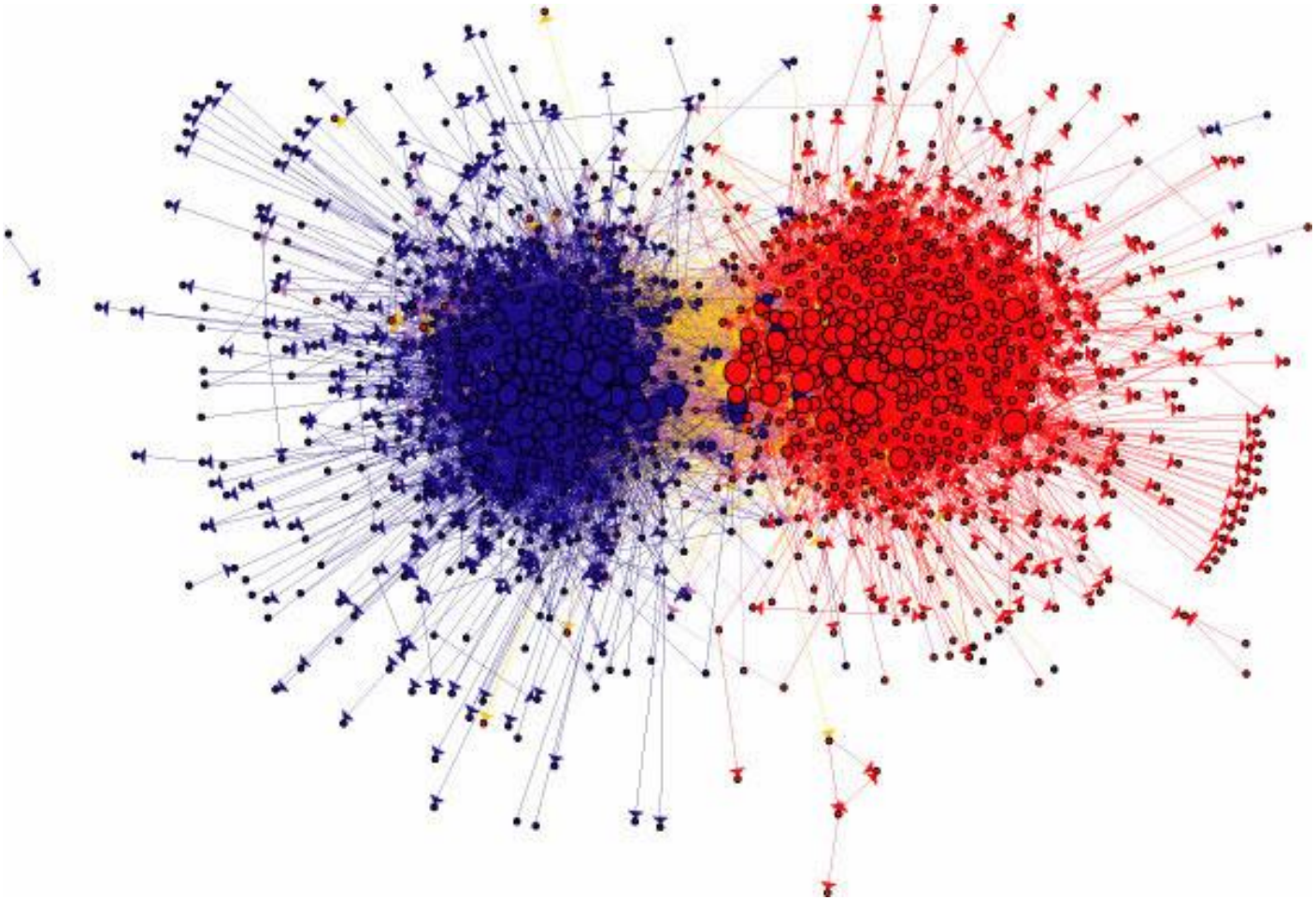
- Can we make up a simple model that exhibits both of the features we've been discussing: many closed triads, but also very short paths?

- In 1998, Duncan Watts and Steve Strogatz argued that such a model follows naturally from a combination of two basic social-network ideas:
  - Weak ties: the links to acquaintances that connect us to parts of the network that would otherwise be far away
    - Weak ties produce very short paths
  - Homophily: the principle that we connect to others who are like ourselves
    - **Homophily creates many closed triads**

Homophily

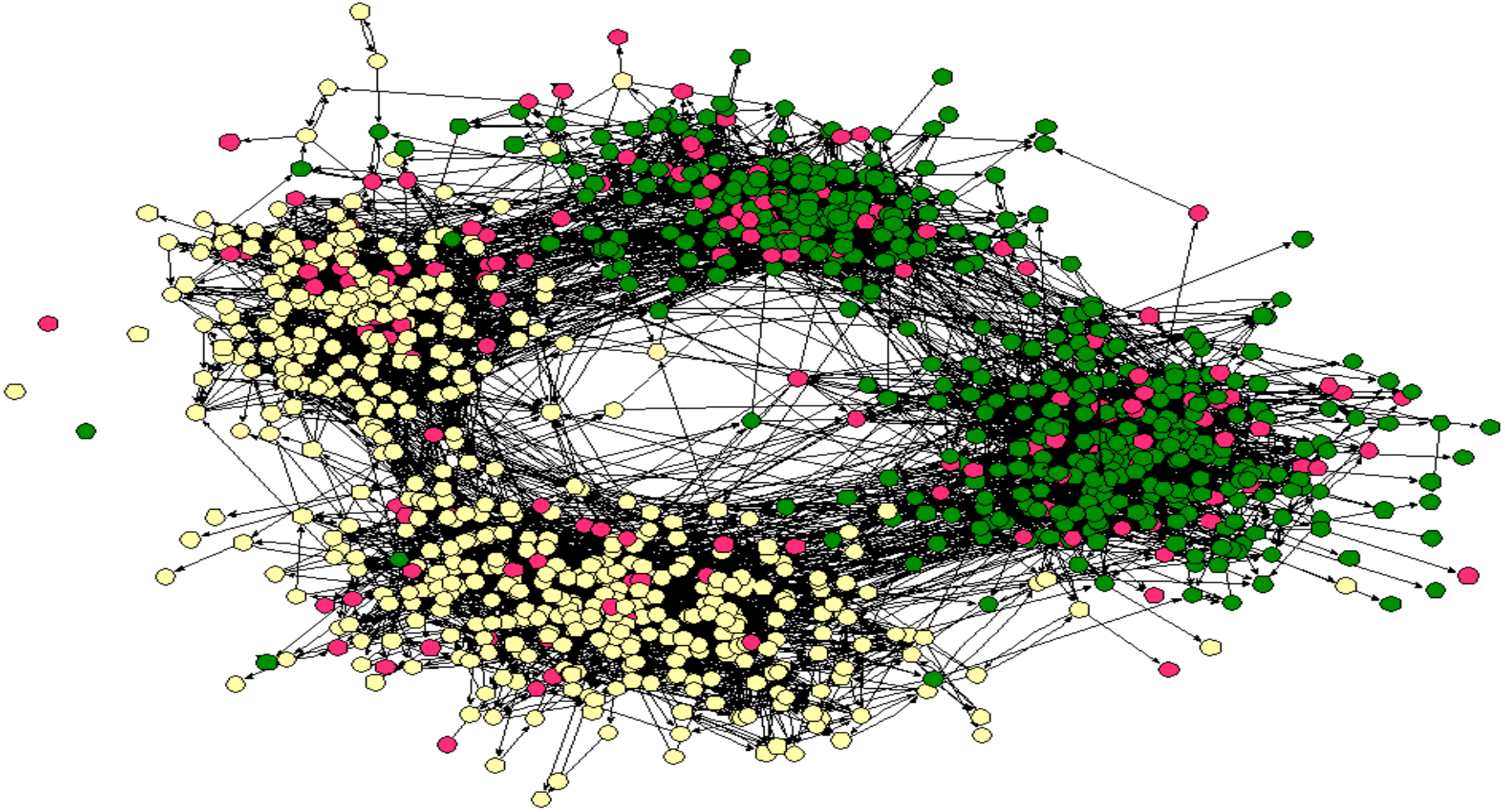
- One of the most basic notions governing the structure of social networks is homophily — the principle that we tend to be similar to our friends.
  - Typically, your friends don't look like a random sample of the underlying population.
    - Viewed collectively, your friends are generally similar to you in terms of race, age, the places they live, their occupations, their levels of affluence, and their interests, beliefs, and opinions.
  - Clearly most of us have specific friendships that cross all these boundaries; but in aggregate, the pervasive fact is that links in a (real-life) social network tend to connect people who are similar to one another.

Homophily



Homophily

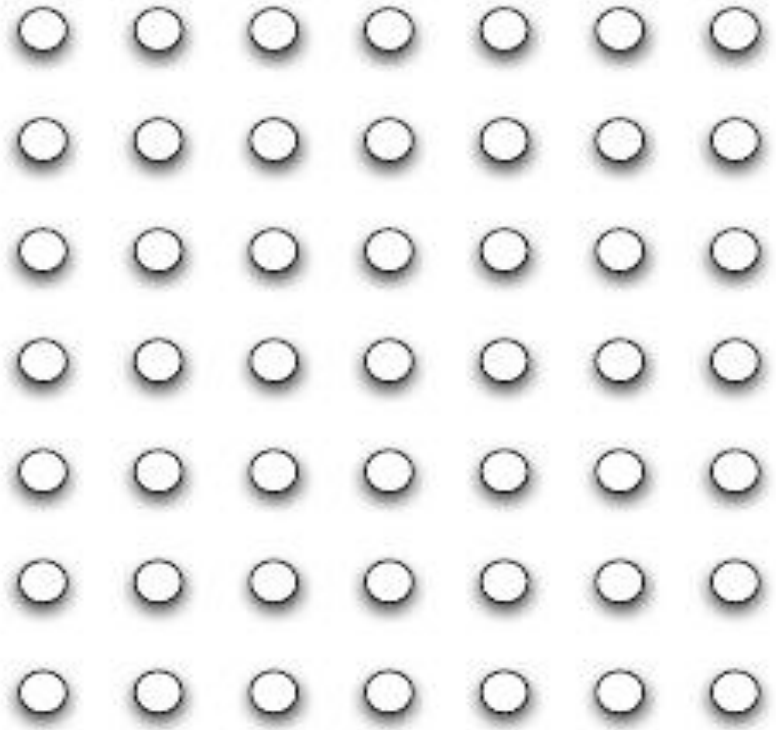
Homophily can produce a division of a social network into densely-connected, homogeneous parts that are weakly connected to each other. In this social network from a town's middle school and high school, two such divisions in the network are apparent: one based on race (with students of different races drawn as differently colored circles), and the other based on friendships in the middle and high schools respectively.

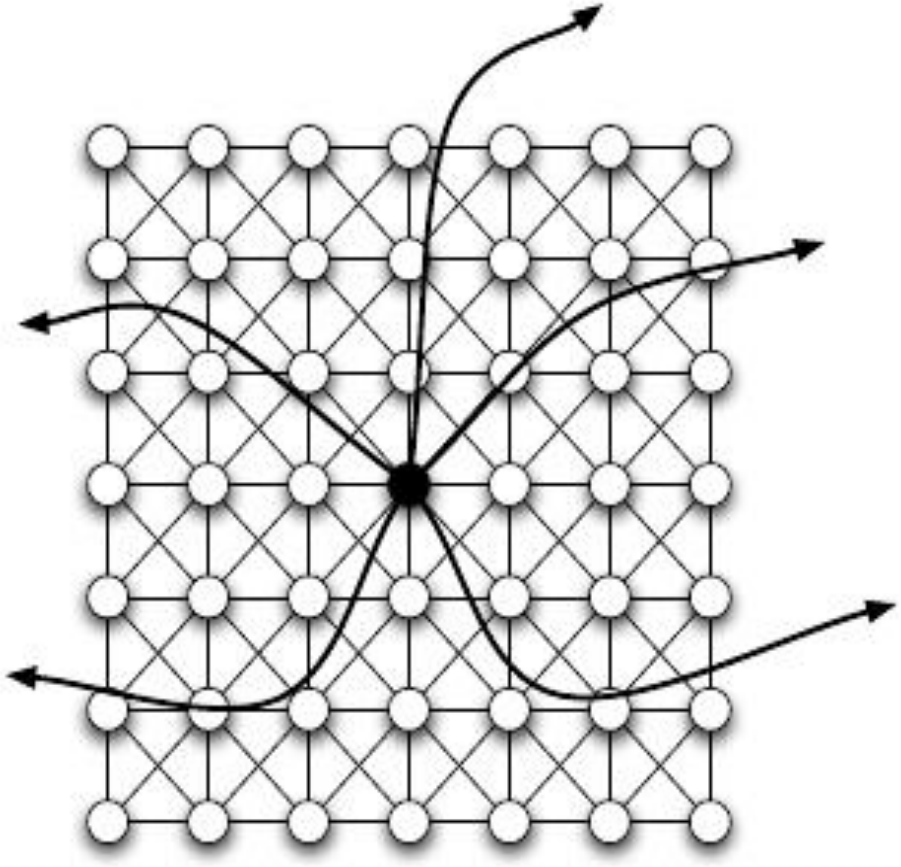


James Moody. Race, school integration, and friendship segregation in america. American Journal of Sociology, 107(3):679-716, November 2001.

The Watts-Strogatz Model

(a) Nodes arranged in a grid

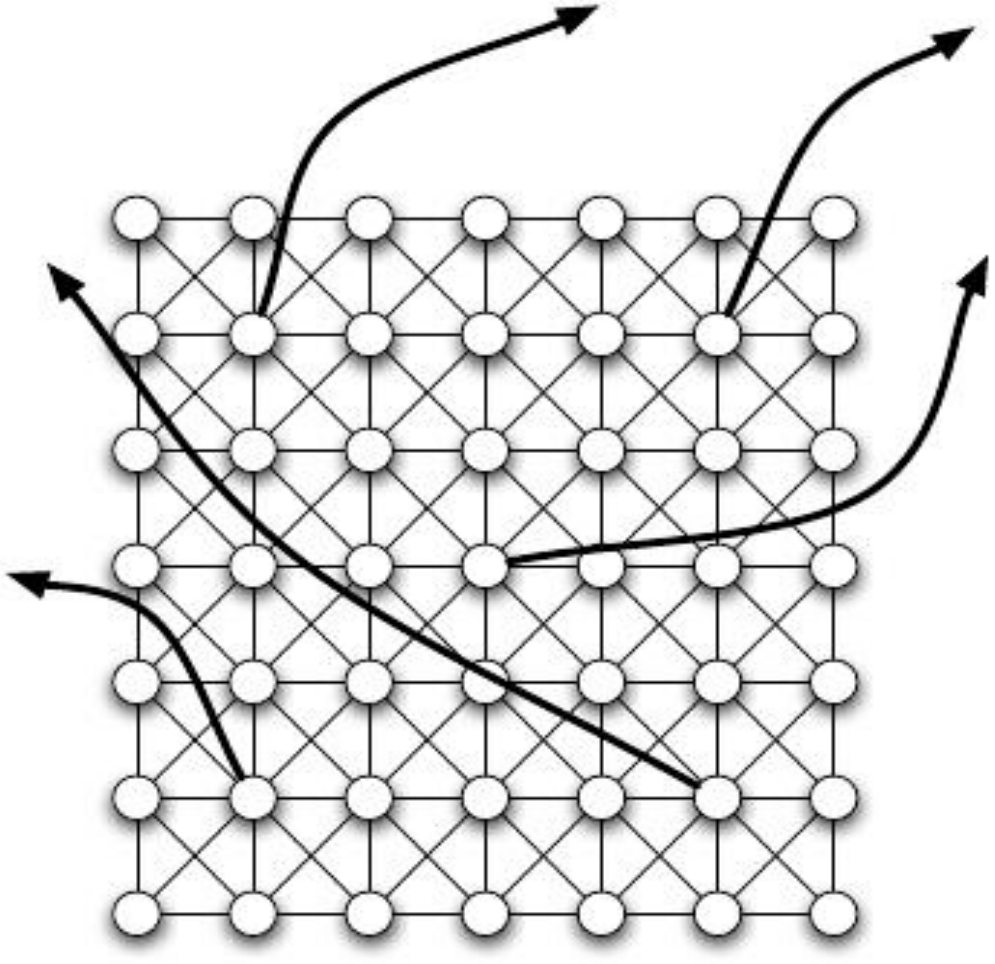




(b) A network built from local structure and random edges

Homophily

The Watts-Strogatz Model

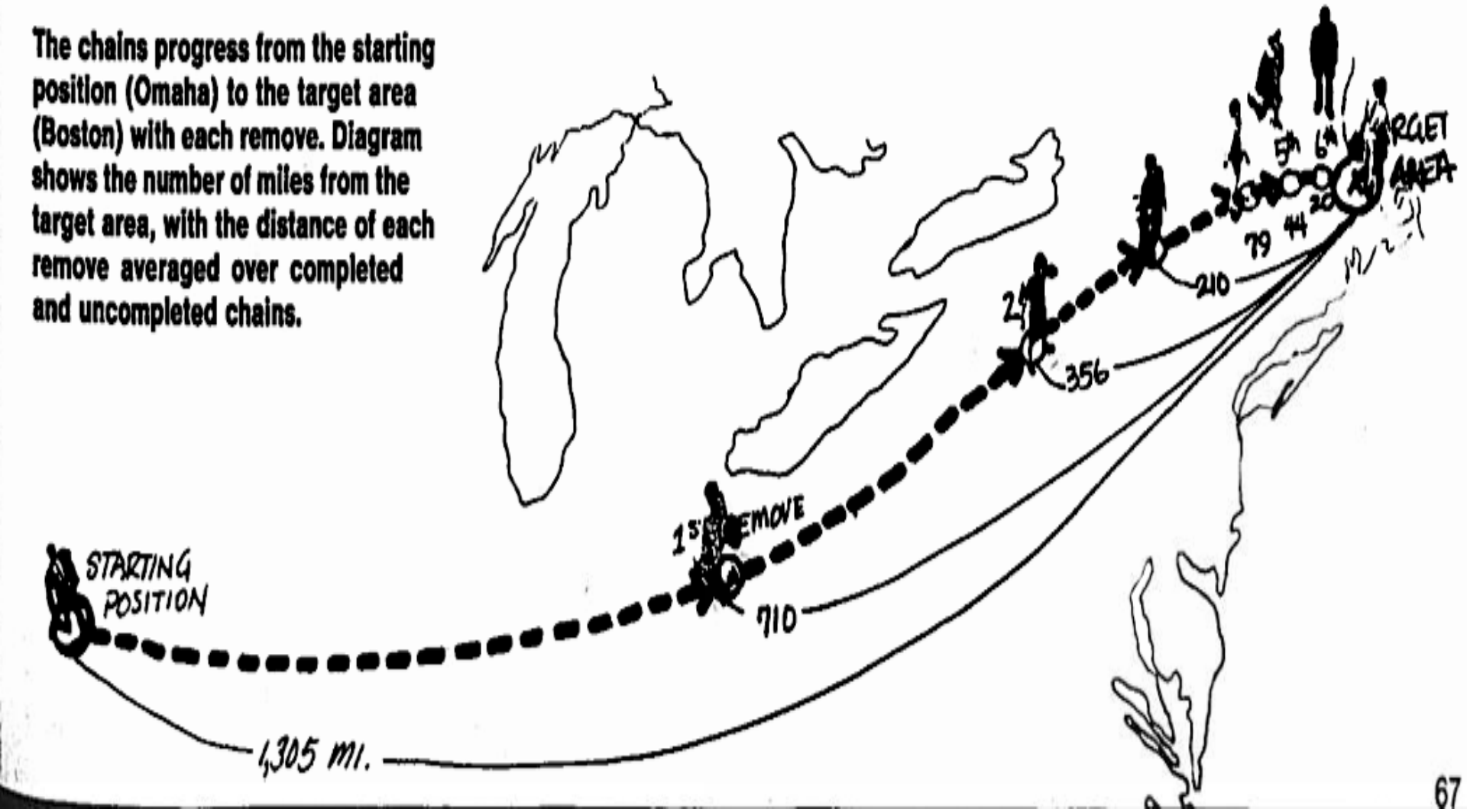


The general conclusions of the Watts-Strogatz model still follow even if only a small fraction of the nodes on the grid each have a ***single*** random link.

Models for Finding Short Paths

- Let's now consider the second basic aspect of the Milgram's small world experiment — the fact that people were actually able to collectively find short paths to the designated target.
  - To really find the shortest path from a starting person to the target, one would have to instruct the starter to forward a letter to all of his or her friends, who in turn should have forwarded the letter to all of their friends, and so forth. (**Breadth-first search**)
    - This "**flooding**" of the network would have reached the target as rapidly as possible, but that was not a feasible option.
  - As a result, Milgram was forced to embark on the much more interesting experiment of constructing paths by "**tunneling**" through the network, with the letter advancing just one person at a time — a process that could well have failed to reach the target, even if a short path existed.

Models for Finding Short Paths



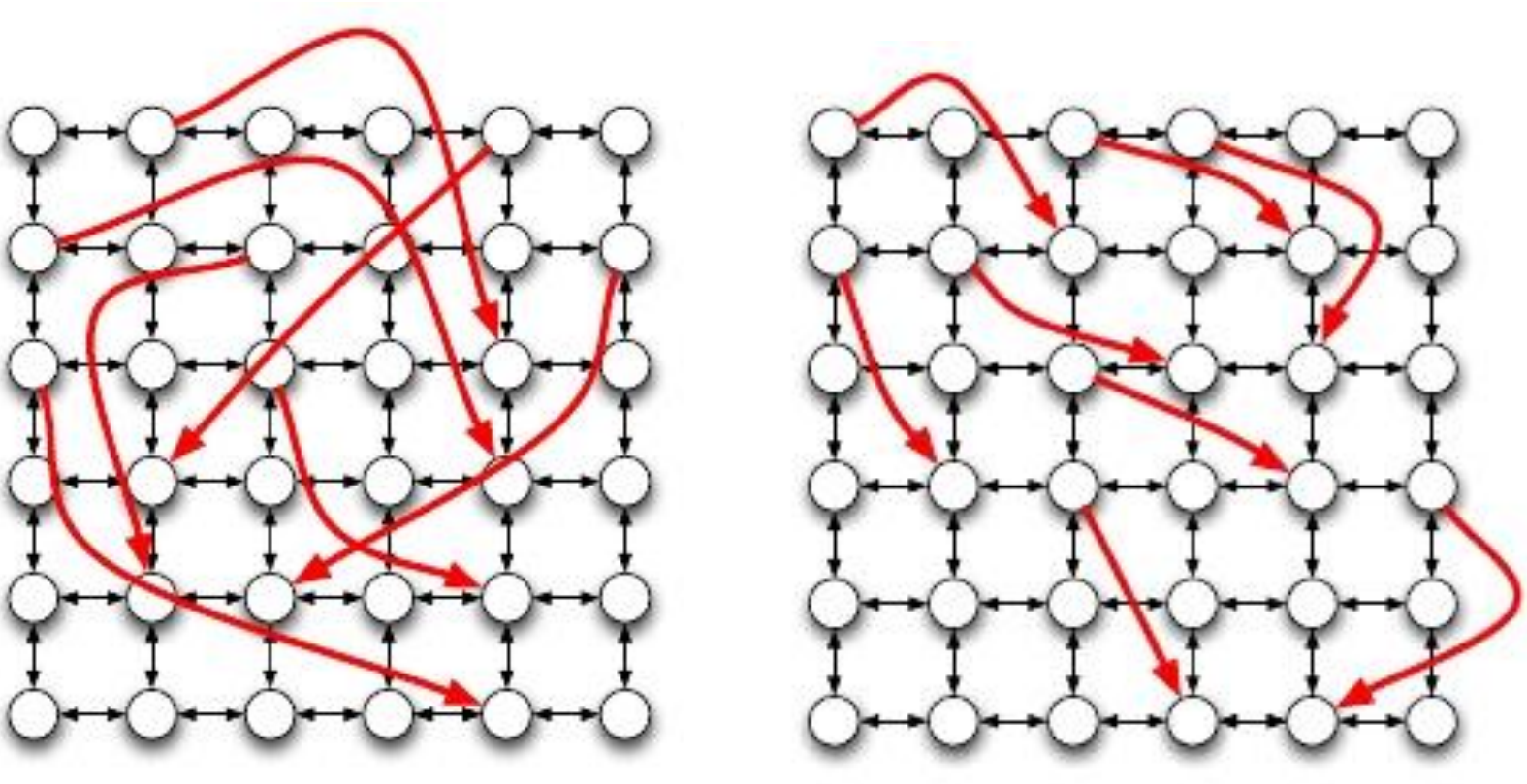
Models for Finding Short Paths

- So the success of the experiment raises fundamental questions about the power of collective search:
  - Even if we all agree that the social network contains short paths, why should the decentralized, 'collective' search process be so effective?
  - Clearly the network contained some type of "**gradient**" that helped participants guide messages toward the target.
- Can we construct a random network in which decentralized search succeeds?
  - As with the Watts-Strogatz model, which sought to provide a simple framework for thinking about short paths in highly clustered networks, this type of search is also something we can try to model

Generalized Watts-Strogatz Model

- We have nodes on a grid as before, and each node still has edges to each other node.
- But now, each of its *k* random edges is generated in a way that it decays with distance, controlled by a clustering exponent ***q*** as follows:
  - For two nodes *v* and *w*, let ***d(v, w)*** denote the number of grid steps between them
  - In generating a random edge out of *v*, we have this edge link to *w* with probability proportional to: $d(v, w)^{-q}$
- We in fact have a different model for each value of *q*
  - The original grid-based model corresponds to $q = 0$, since the links are chosen uniformly at random
  - Varying *q* is like turning a knob that controls how uniform the random links are.

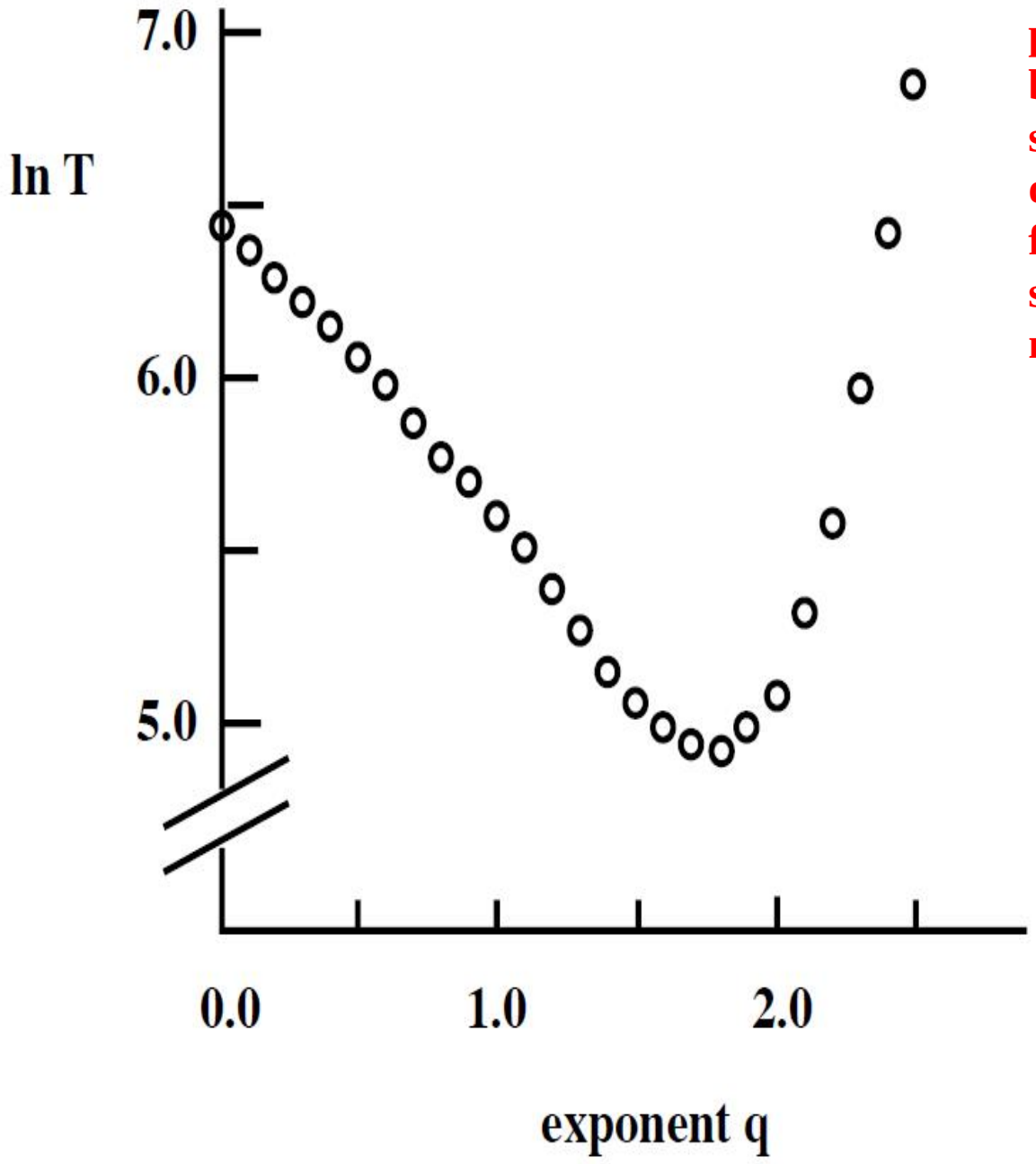
Generalized Watts-Strogatz Model



(a) A small clustering exponent, *q*

(b) A large clustering exponent, q

(For in-depth analysis, see pp.554-564, **Network, Crowds and Markets**, Chapter 20)



This chart shows the performance of a basic decentralized search method across different values of *q*, for a network of several hundred million nodes.