SOCIAL MEDIA AND NETWORKS

Alan Mislove

College of Computer and Information Science

Northeastern University

NetSci 2010 International School

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, Instant Messaging

Social news media

- Blogs, iReport

Online worlds

- World of WarCraft, Second Life

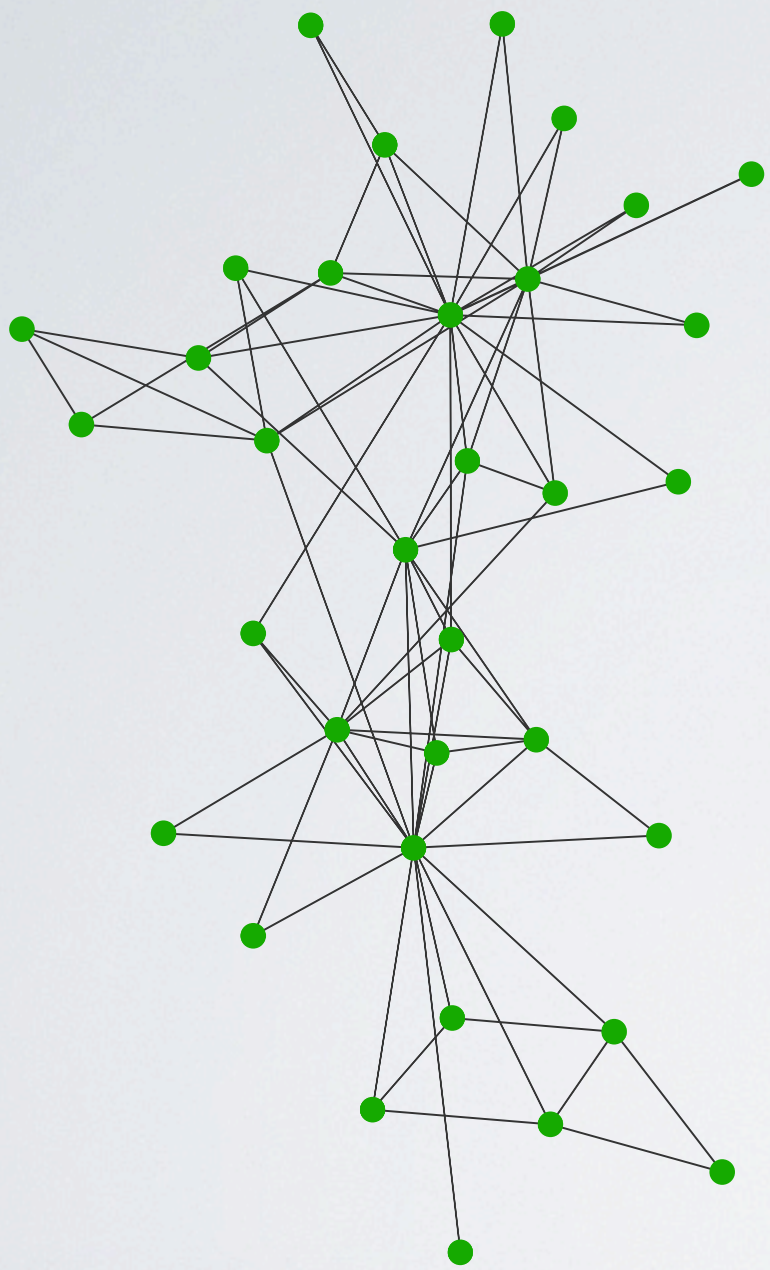
Why is social media interesting?

Two reasons (to me):

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

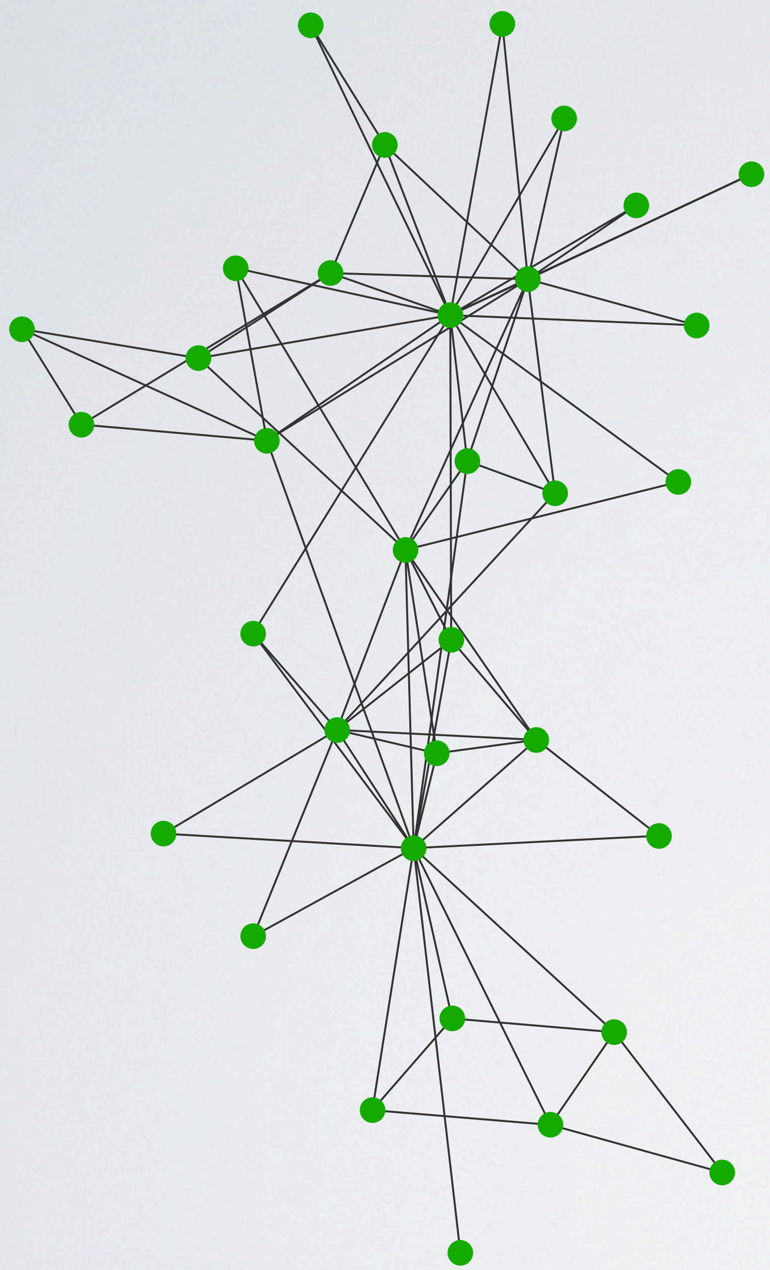
2. Relate information and people
   - Online social networks now content-sharing systems
   - Can attach reputation of users to content

1. Observe social interactions



Anyone recognize this network?

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

- 50% of Facebook users online on given day
- 500B people-minutes spent on Facebook each month
- Every interaction recorded

What scale has been studied before?

1977: 34 people in a Karate club

[Zachary, J. Anth. Res. 1972]

2003: 436 people using a corporate email system

[Adamic and Adar, Social Networks 2003]

2006: 43,553 people using a university email system

[Kossinets and Watts, Science 2006]

2007: 4,400,000 people using an online blogging service

[Backstrom et al., KDD 2007]

2009: 240,000,000 people using an instant messaging service

[Leskovec and Horvitz, WWW 2008]

(stats and slides borrowed from Jure Leskovec)

The curse of scale

Scale is both a blessing and a curse

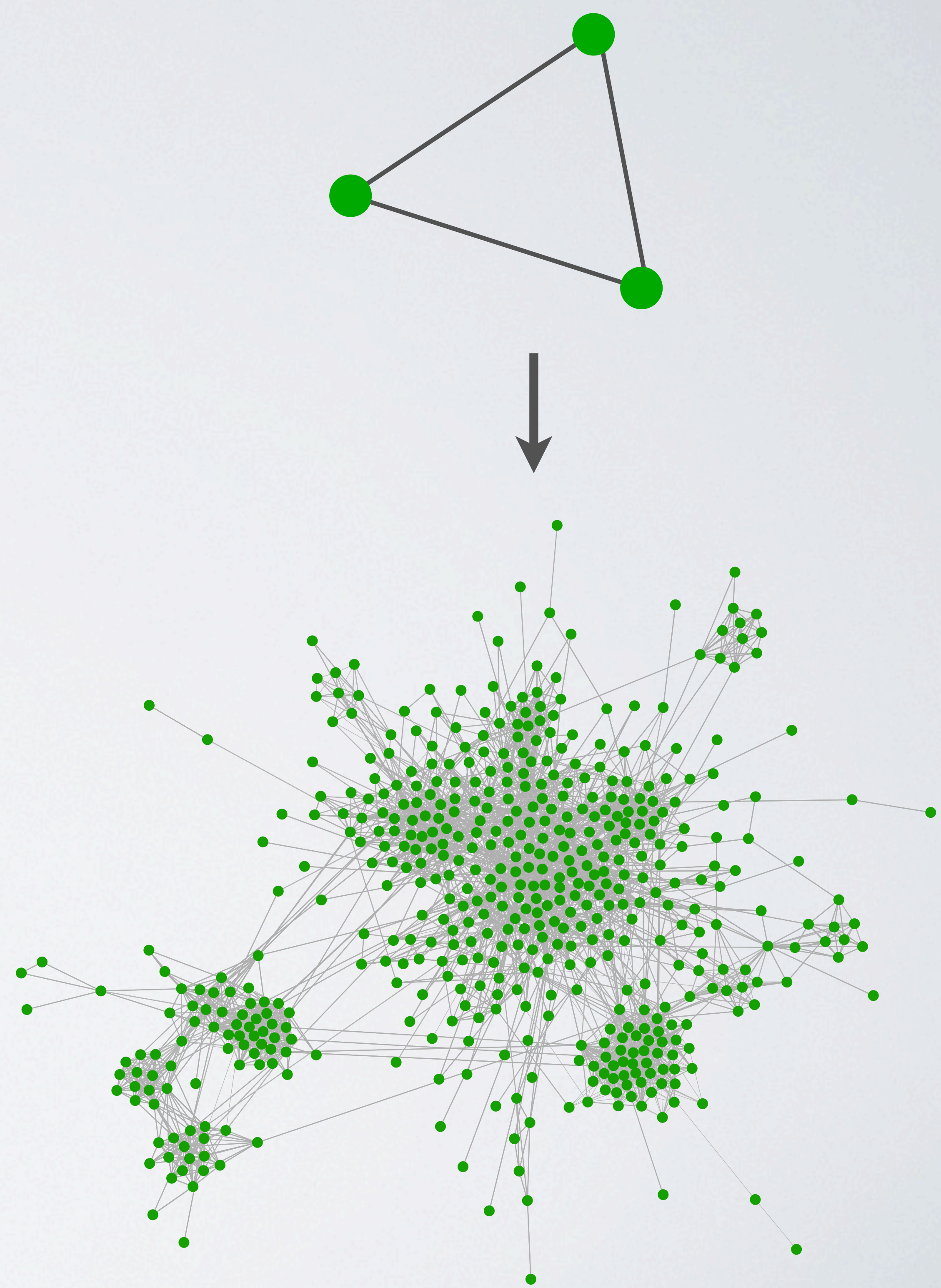
Blessing

- 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...
- MySpace (275 M), Facebook (300 M)

Growing exponentially

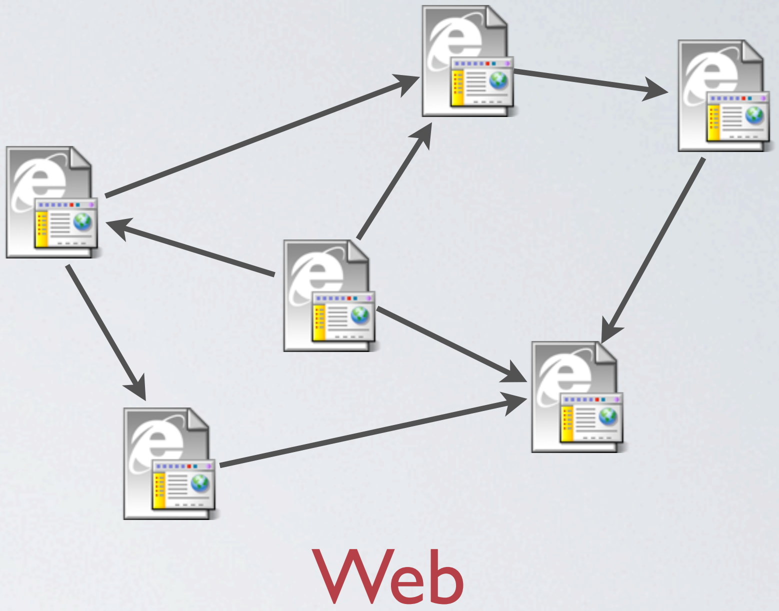
Incredible amounts of content being shared

- Facebook (850 M photos/month)
- YouTube (24 hours of video/min)

A new way of organizing information

Web organized with content-content links

- Link structure exploited (e.g., PageRank)



Social media organized using

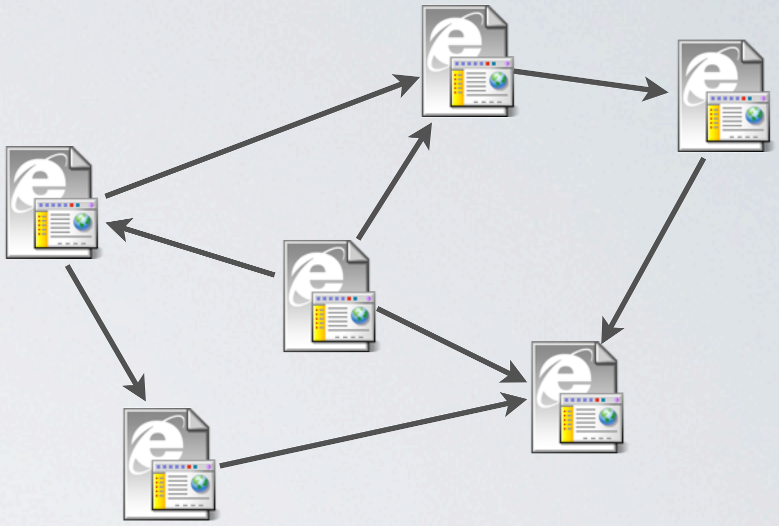
- User-user links (social network)
- User-content links (favorites, etc)

New platform for information sharing

A new way of organizing information

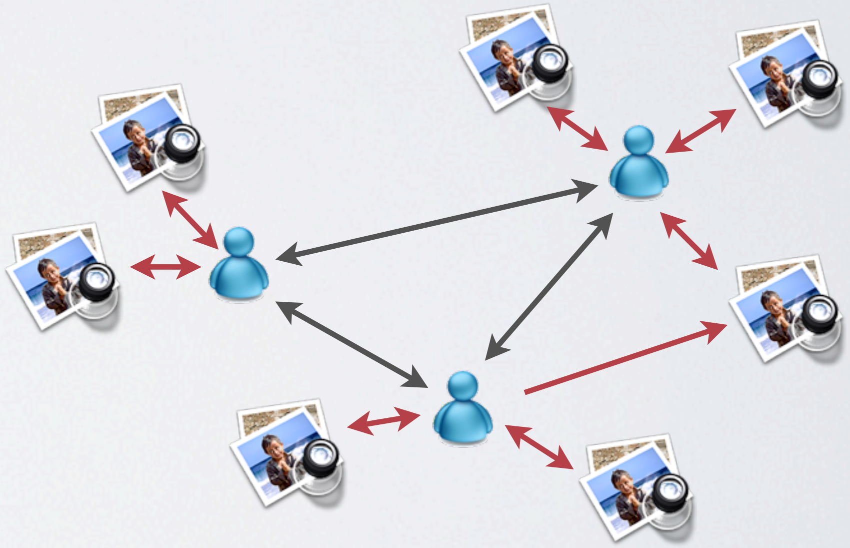
Web organized with content-content links

- Link structure exploited (e.g., PageRank)



Social media organized using

- User-user links (social network)
- User-content links (favorites, etc)



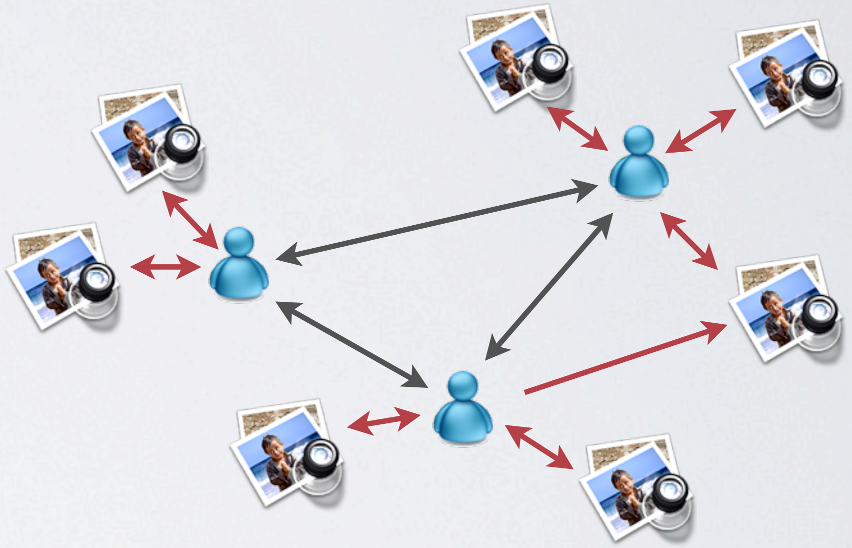
New platform for information sharing

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

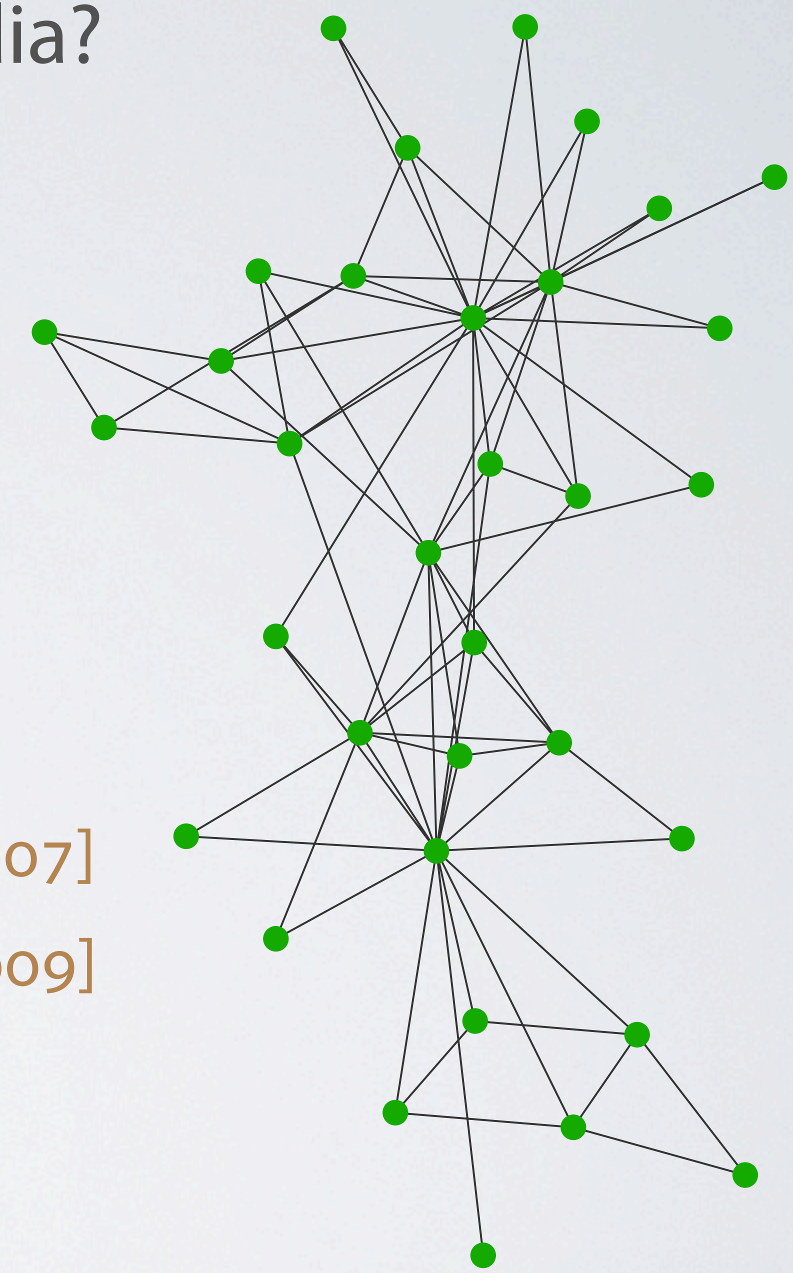
Obesity [Fowler and Christakis, NE J. Med. 2007]

Altruism [Fowler et al., Econ. Let. 2009]

Example: Zachary's Karate Club

Can predict behavior with network view

[Zachary, J. Anth. Res. 1972]



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

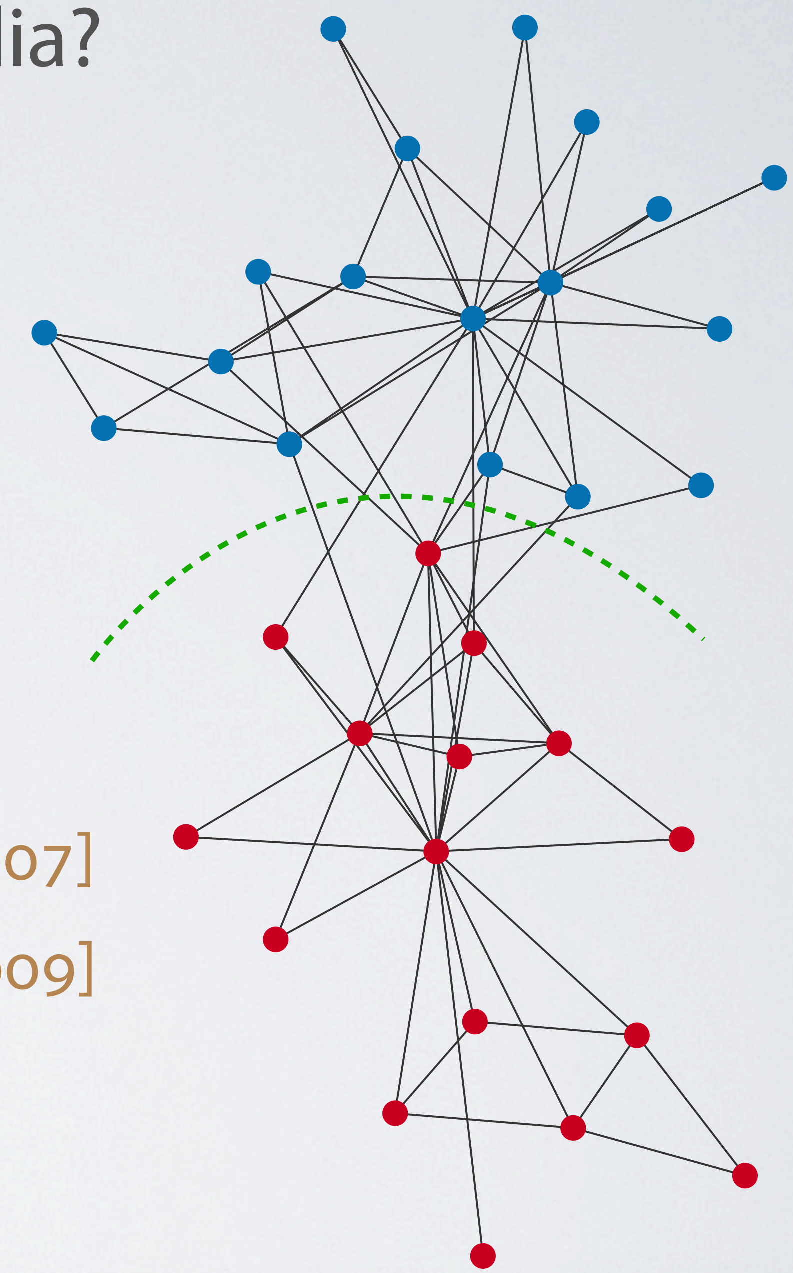
Obesity [Fowler and Christakis, NE J. Med. 2007]

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?

This lecture

- Basically, discuss things I find interesting
  - By no means an exhaustive list
- To understand social media, need to understand social interactions
  - Thus, will cover some sociology, anthropology, ...
- Examine how users are interacting on social media sites
  - And explore what these sites can tell us
- Slides "borrowed" from many places
  - Jure Leskovec in particular

Outline

Four parts:

1 Primer on social sciences

2 Measuring social media

3 Leveraging social media

4 Open questions

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!*

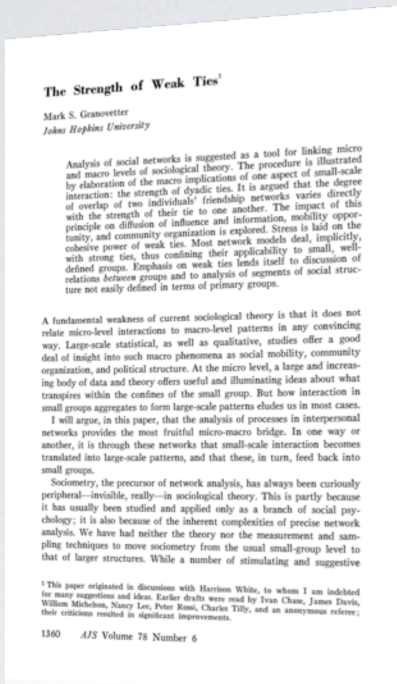
PART I

Primer on social sciences

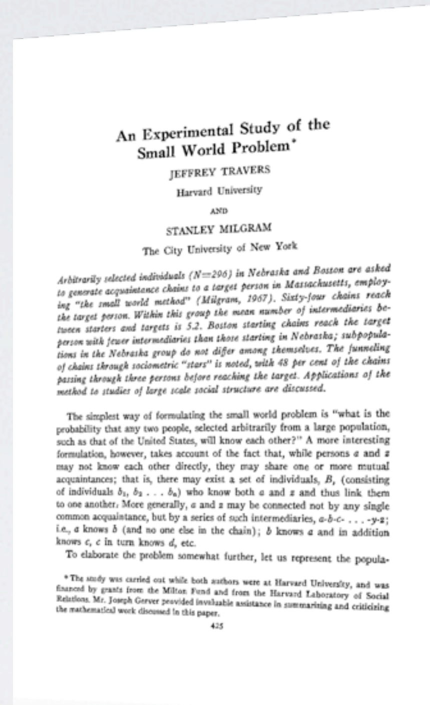
- or -

What have people studied this before?

Three classic papers


The Strength of Weak Ties

Mark S. Granovetter
Johns Hopkins University


An Experimental Study of the Small World Problem

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


R. I. M. Dunbar

Neocortex size as a constraint on group size in primates

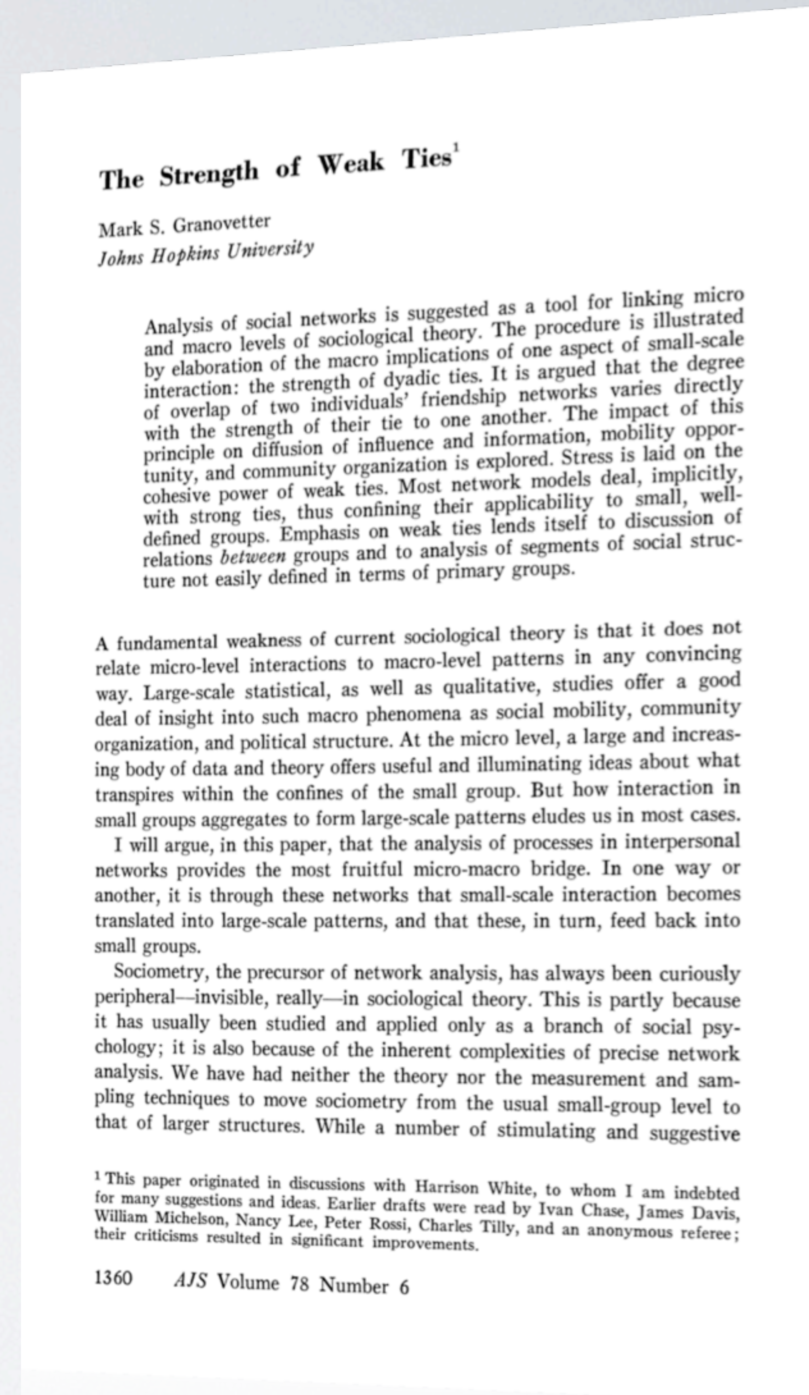
Introduction

- 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.

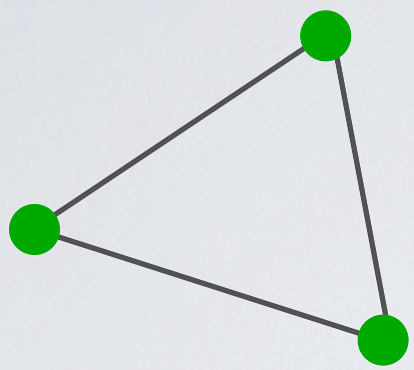
A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, 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

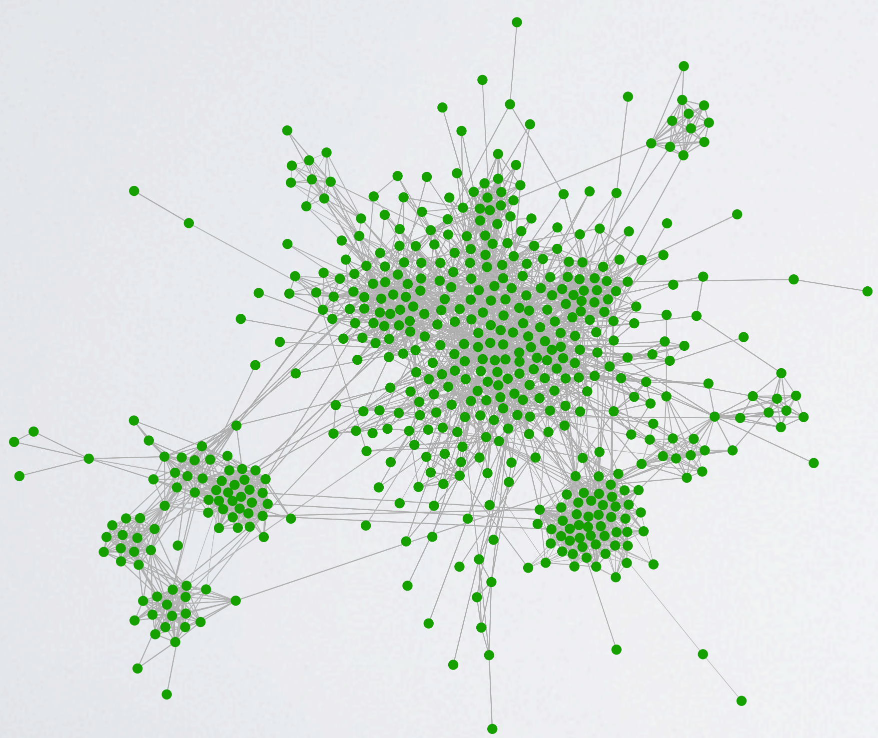
[1] This paper originated in discussions with Harrison White, to whom I am indebted for many suggestions and ideas. Earlier drafts were read by Ivan Chase, James Davis, William Michelson, Nancy Lee, Peter Rossi, Charles Tilly, and an anonymous referee; their criticisms resulted in significant improvements.

"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

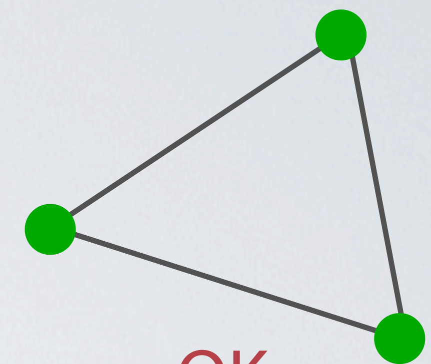
But, mapping not understood

- How do large-scale patterns emerge?
- ...certain analogies to physics...

Granovetter’s idea

Construct simple model:

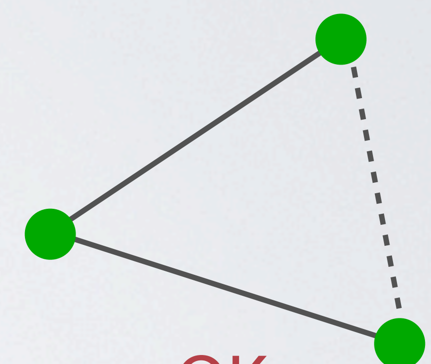
If two people have a common strong tie, they must have a tie between each other



OK

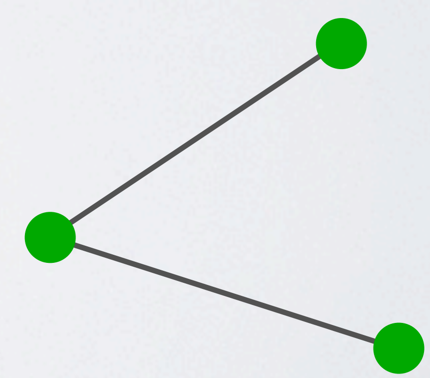
Matches intuition from real world

If you have two close friends, they (at least) know each other



OK

What are the implications of this model?



Forbidden

Bridges

Social networks can be divided into communities

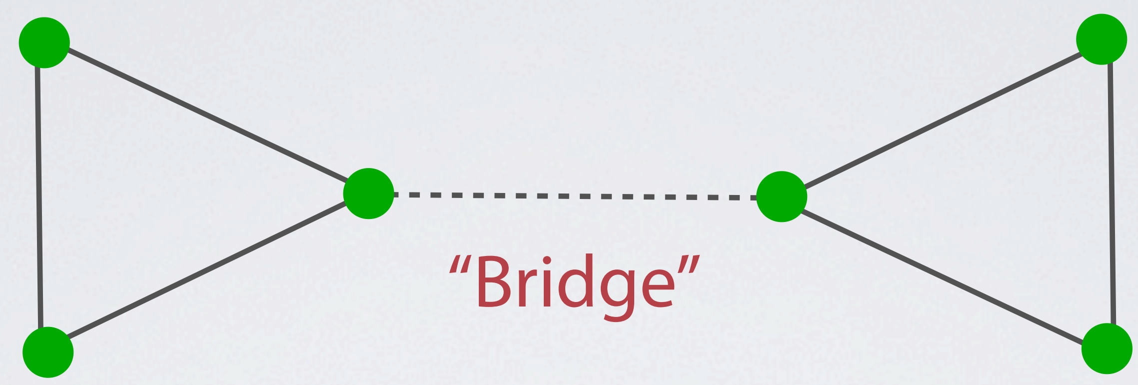
- Clubs, schools, employers, ...

Define a *bridge* as a link that is the only path between two users

Claim: With Granovetter's assumption, bridges must be weak

- Why?

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 the really necessary for conveying information

An Experimental Study of the Small World Problem

by Jeffery Travers and Stanley Milgram

[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 $b_1, b_2 \ldots b_n$) 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.

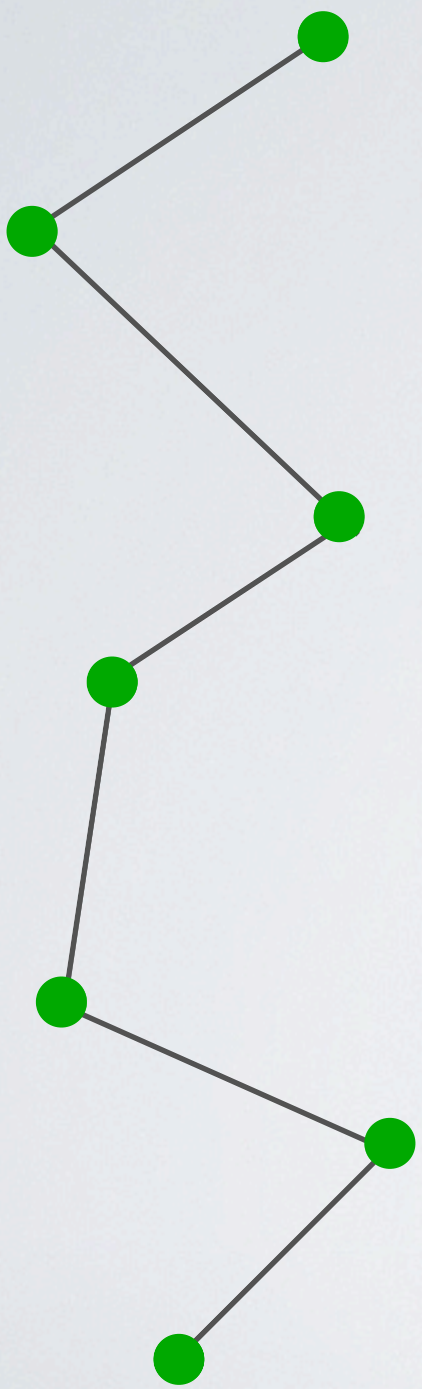
To elaborate the problem somewhat further, let us represent the popula-

* 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.

Six degrees of Kevin Bacon

- *Six degrees of separation* now common saying
  - Popularized by this study
- At the time, sociologists had no idea of shortest path
- Assume 200M people, each with ~100 friends
  - Expect 2-3 intermediaries
  - But does not consider network structure
- High clustering increases path lengths
  - What is the actual value?
  - How to measure?

Six degrees of Kevin Bacon



- *Six degrees of separation* now common saying
  - Popularized by this study
- At the time, sociologists had no idea of shortest path
- Assume 200M people, each with ~100 friends
  - Expect 2-3 intermediaries
  - But does not consider network structure
- High clustering increases path lengths
  - What is the actual value?
  - How to measure?

Procedure

Selected 296 people in Nebraska and Boston

- Mailed a packet containing instructions

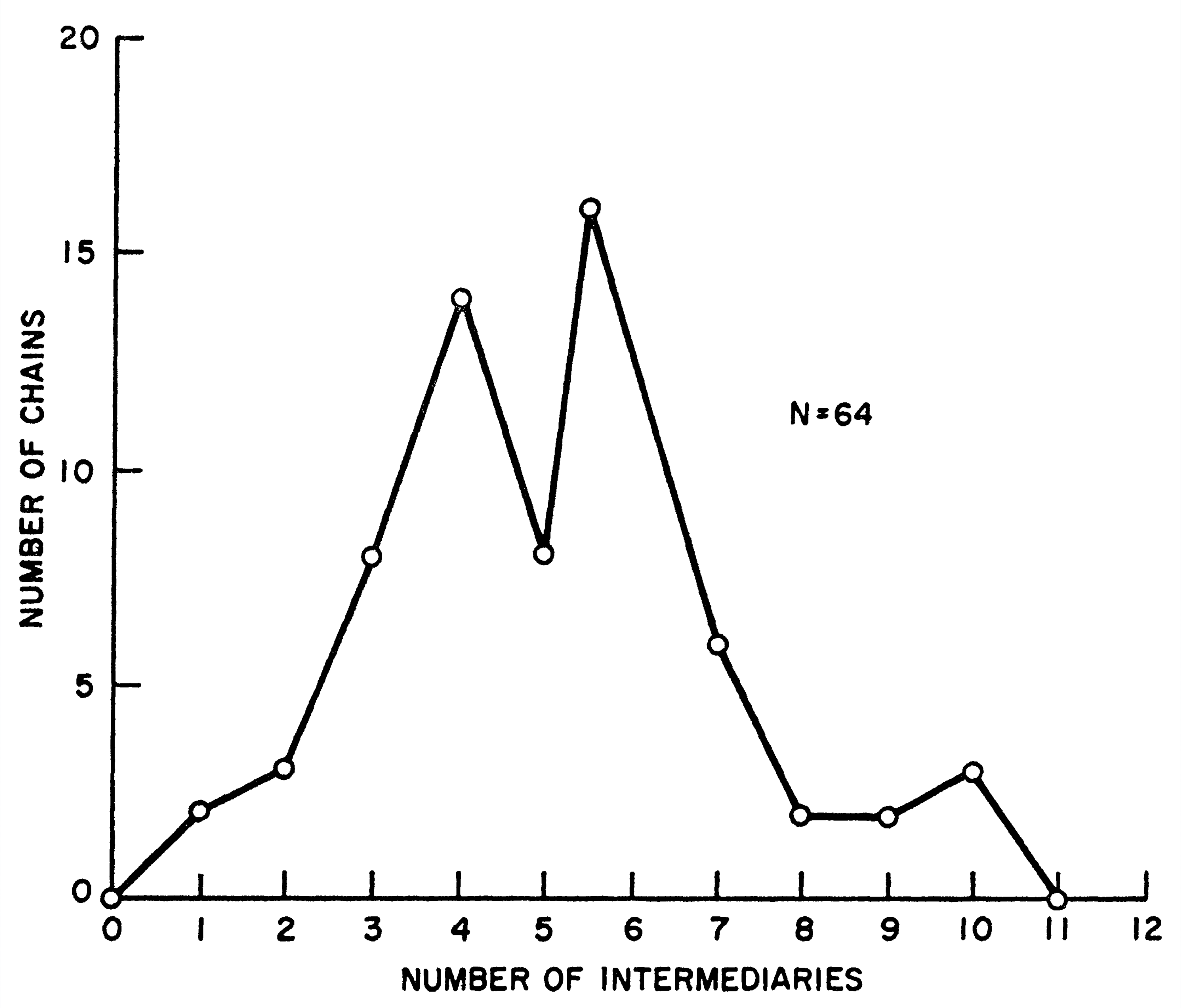
Packet specified a destination person

- 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

- 40% of chains coalesced into 2 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 on Group Size in Primates

by Robin I. M. Dunbar

[Journal of Human Evolution, vol. 22. 1992]

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

In hominids, represents 80% of brain by volume

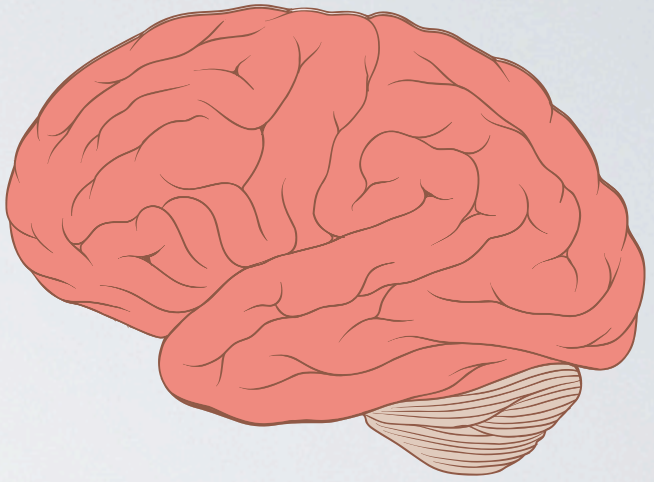
Theory: large brain size due to "social" nature of primates

- Measure "social" level by looking at typical group size
- If true, then brain size should correlate with being "social"

Neocortex

Part of the brain of mammals, involved in

- Sensory perception
- Motor commands
- Spatial reasoning
- Thought and language
- Social interactions



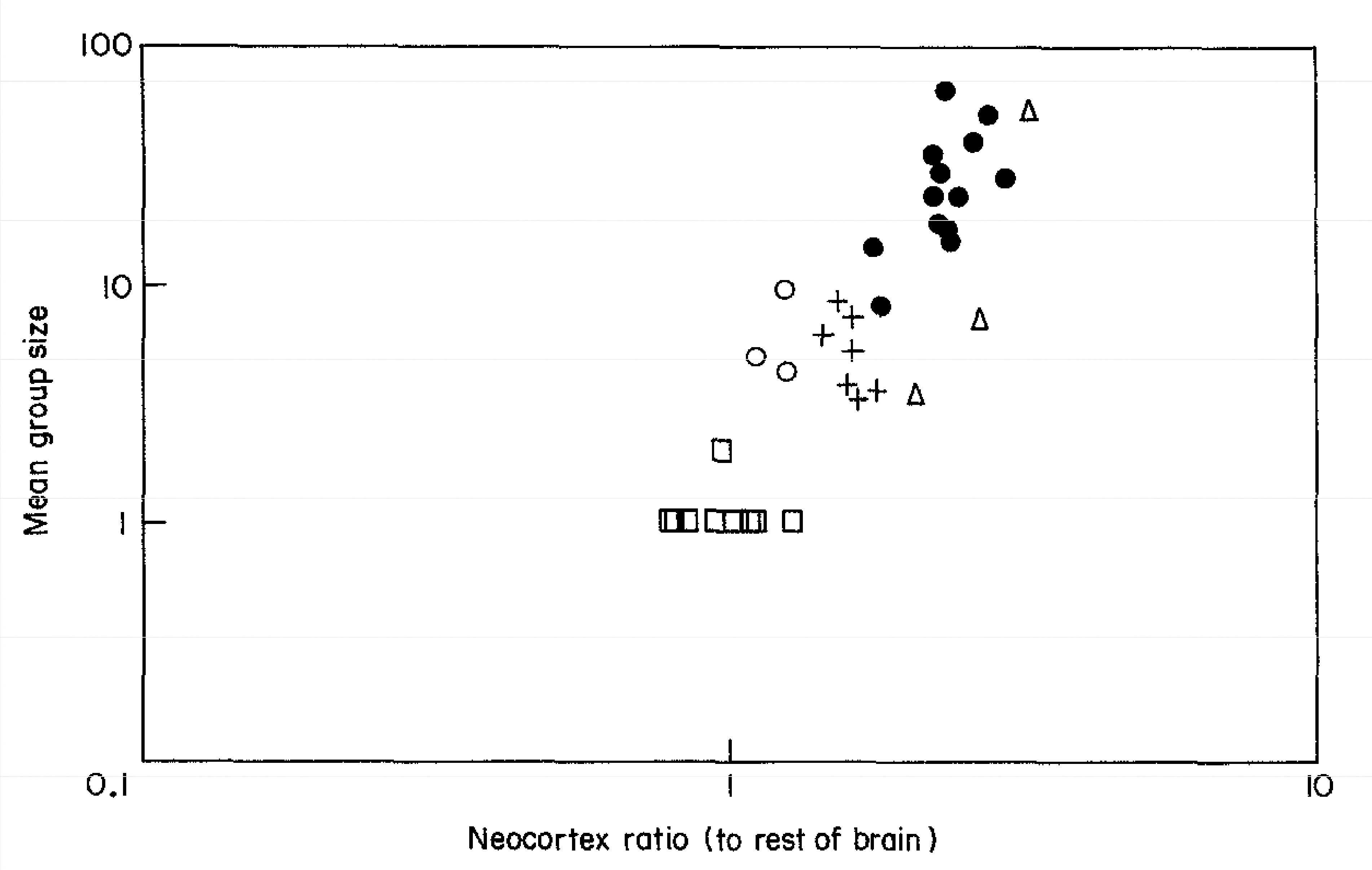
Neocortex

In hominids, represents 80% of brain by volume

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

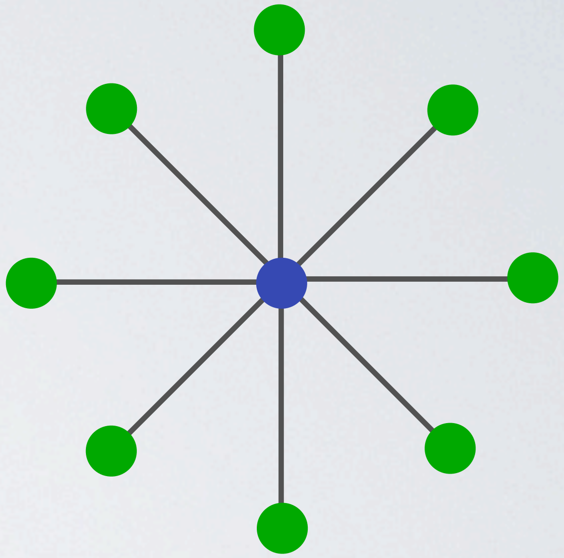
Implications

Each individual can only maintain so many relationships

- Bounded by brain size

Not just *number* of relationships, it's the pairs (dyads)

- 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!

- Zachary: 34 people, Milgram: 64 chains, Dunbar: 43 people

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?

PART II

Measuring social media

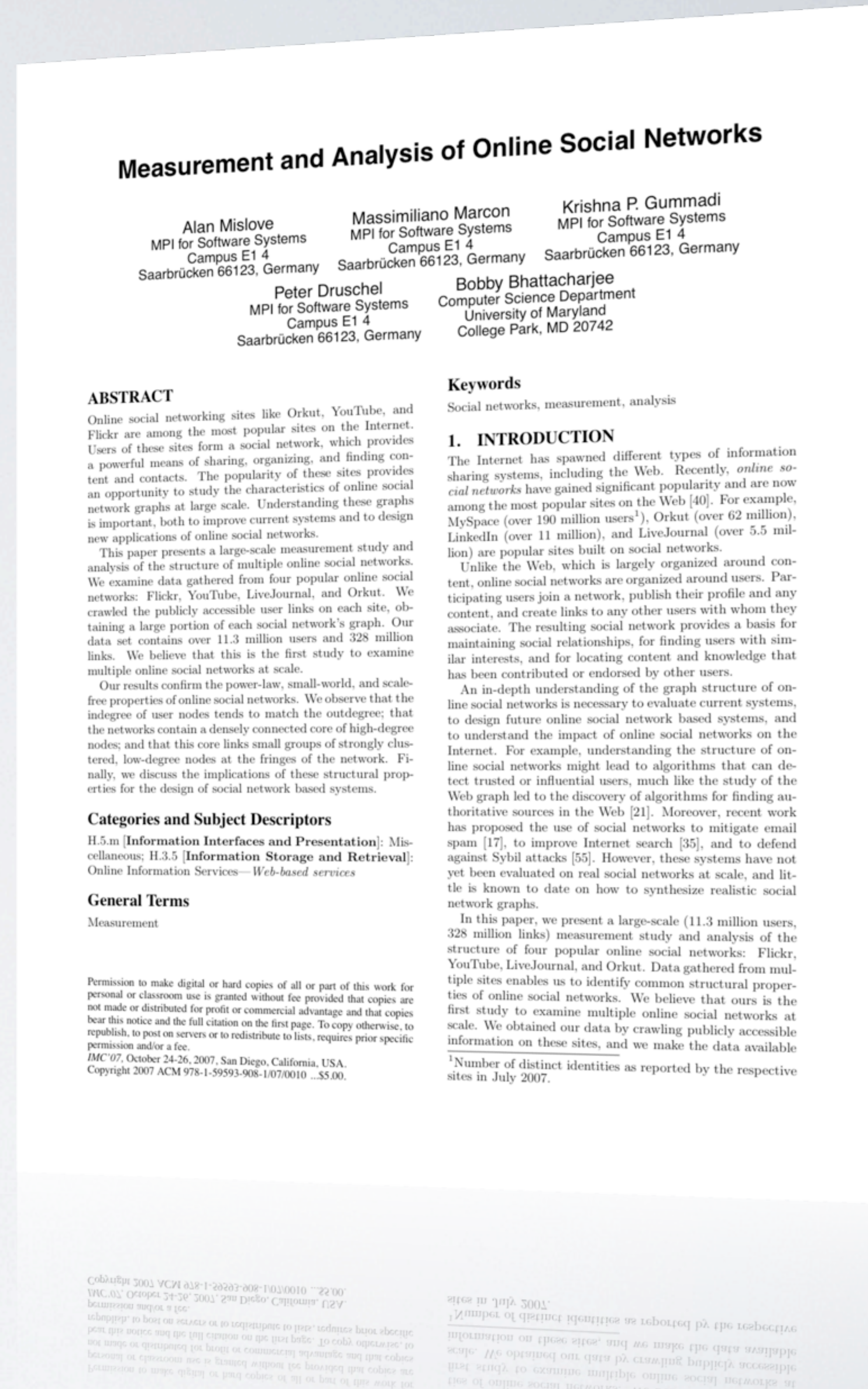
- or -

What does Facebook look like?

Measurement and Analysis of Online Social Networks

by Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee

[Proceedings of IMC 2007]



Measurement and Analysis of Online Social Networks

Alan Mislove
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Massimiliano Marcon
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Krishna P. Gummadi
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Peter Druschel
MPI for Software Systems
Campus E1 4
Saarbrücken 66123, Germany

Bobby Bhattacharjee
Computer Science Department
University of Maryland
College Park, MD 20742

ABSTRACT

Online social networking sites like Orkut, YouTube, and Flickr are among the most popular sites on the Internet. Users of these sites form a social network, which provides a powerful means of sharing, organizing, and finding content and contacts. The popularity of these sites provides an opportunity to study the characteristics of online social network graphs at large scale. Understanding these graphs is important, both to improve current systems and to design new applications of online social networks.

This paper presents a large-scale measurement study and analysis of the structure of multiple online social networks. We examine data gathered from four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. We crawled the publicly accessible user links on each site, obtaining a large portion of each social network's graph. Our data set contains over 11.3 million users and 328 million links. We believe that this is the first study to examine multiple online social networks at scale.

Our results confirm the power-law, small-world, and scale-free properties of online social networks. We observe that the indegree of user nodes tends to match the outdegree; that the networks contain a densely connected core of high-degree nodes; and that this core links small groups of strongly clustered, low-degree nodes at the fringes of the network. Finally, we discuss the implications of these structural properties for the design of social network based systems.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation**]: Miscellaneous; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Measurement

Keywords

Social networks, measurement, analysis

1. INTRODUCTION

The Internet has spawned different types of information sharing systems, including the Web. Recently, *online social networks* have gained significant popularity and are now among the most popular sites on the Web [40]. For example, MySpace (over 190 million users[1]), Orkut (over 62 million), LinkedIn (over 11 million), and LiveJournal (over 5.5 million) are popular sites built on social networks.

Unlike the Web, which is largely organized around content, online social networks are organized around users. Participating users join a network, publish their profile and any content, and create links to any other users with whom they associate. The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users.

An in-depth understanding of the graph structure of online social networks is necessary to evaluate current systems, to design future online social network based systems, and to understand the impact of online social networks on the Internet. For example, understanding the structure of online social networks might lead to algorithms that can detect trusted or influential users, much like the study of the Web graph led to the discovery of algorithms for finding authoritative sources in the Web [21]. Moreover, recent work has proposed the use of social networks to mitigate email spam [17], to improve Internet search [35], and to defend against Sybil attacks [55]. However, these systems have not yet been evaluated on real social networks at scale, and little is known to date on how to synthesize realistic social network graphs.

In this paper, we present a large-scale (11.3 million users, 328 million links) measurement study and analysis of the structure of four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. Data gathered from multiple sites enables us to identify common structural properties of online social networks. We believe that ours is the first study to examine multiple online social networks at scale. We obtained our data by crawling publicly accessible information on these sites, and we make the data available

[1] Number of distinct identities as reported by the respective sites in July 2007.

Myriad of social networking sites

Social media captures some notion of friendship

- How hard is it to be Facebook 'friends'?

'Friend'-ship has different implications

- Flickr: bookmark
- LinkedIn: send messages
- Facebook: view (some) content

Question: Do mechanisms and policies make social networks look different?

Comparing multiple sites

Measurement study of the structure of multiple online social networks

- 11 M users, 328 M links

Data from four diverse online social networks

- Flickr: photo sharing
- LiveJournal: blogging site
- Orkut: social networking site
- YouTube: video sharing

Goals are two-fold:

- Measure online social networks at scale
- Understand static structural properties

Measuring social networks

Sites reluctant to give out data

- Cannot enumerate user list
- Instead, performed crawls of user graph

Picked known seed user

- Crawled all of his friends
- Added new users to list

Continued until all known users crawled

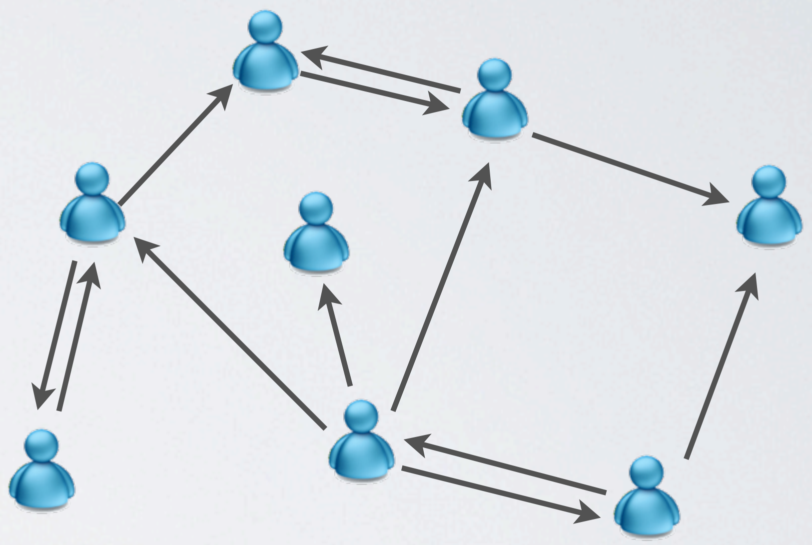
Effectively performed a BFS of graph

- Reachable from seed user

Measuring social networks

Sites reluctant to give out data

- Cannot enumerate user list
- Instead, performed crawls of user graph



Picked known seed user

- Crawled all of his friends
- Added new users to list

Continued until all known users crawled

Effectively performed a BFS of graph

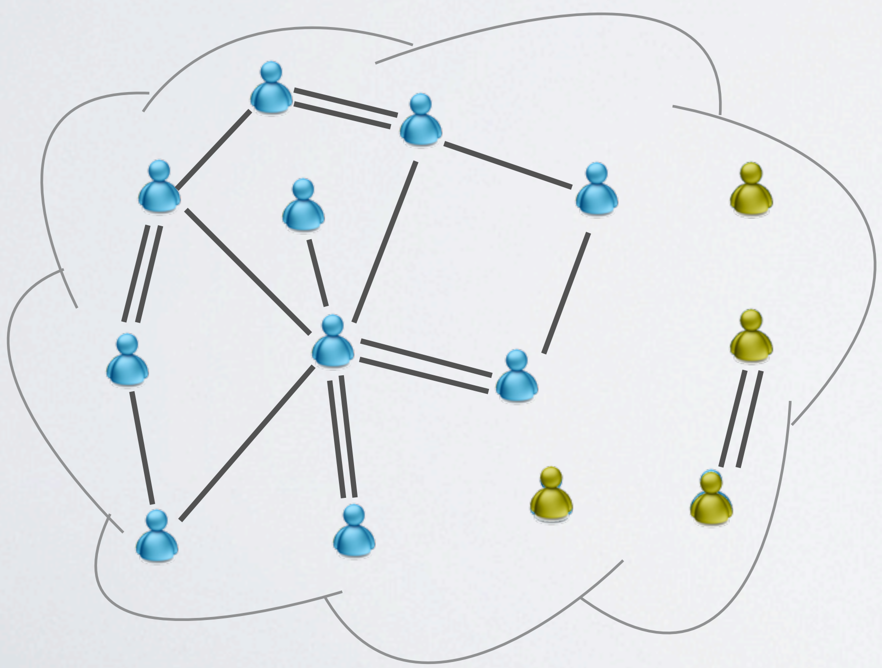
- Reachable from seed user

Challenges

Obtaining data using crawling presents unique challenges

Crawling quickly

- Underlying social networks changing rapidly
- Consistent snapshot hard to get
- Need to complete the crawl quickly



Crawling completely

- Social networks aren't necessarily connected
- Some users have no links, or small clusters
- Need to estimate the crawl coverage

Data collected

	Flickr	LiveJournal	Orkut	YouTube
Number of Users				
Avg. Friends per User				

Able to crawl large portion of networks

Node degrees vary by orders of magnitude

- However, networks share many key properties

To ground analysis, will compare to Web [Broder et al., INFOCOM'99]

Data collected

	Flickr	LiveJournal	Orkut	YouTube
Number of Users	1.8 M	5.2 M	3.0 M	1.1 M
Avg. Friends per User				

Able to crawl large portion of networks

Node degrees vary by orders of magnitude

- However, networks share many key properties

To ground analysis, will compare to Web [Broder et al., INFOCOM'99]

Data collected

	Flickr	LiveJournal	Orkut	YouTube
Number of Users	1.8 M	5.2 M	3.0 M	1.1 M
Avg. Friends per User	12.2	16.9	106.1	4.2

Able to crawl large portion of networks

Node degrees vary by orders of magnitude

However, networks share many key properties

To ground analysis, will compare to Web [Broder et al., INFOCOM'99]

Are online social networks power-law?

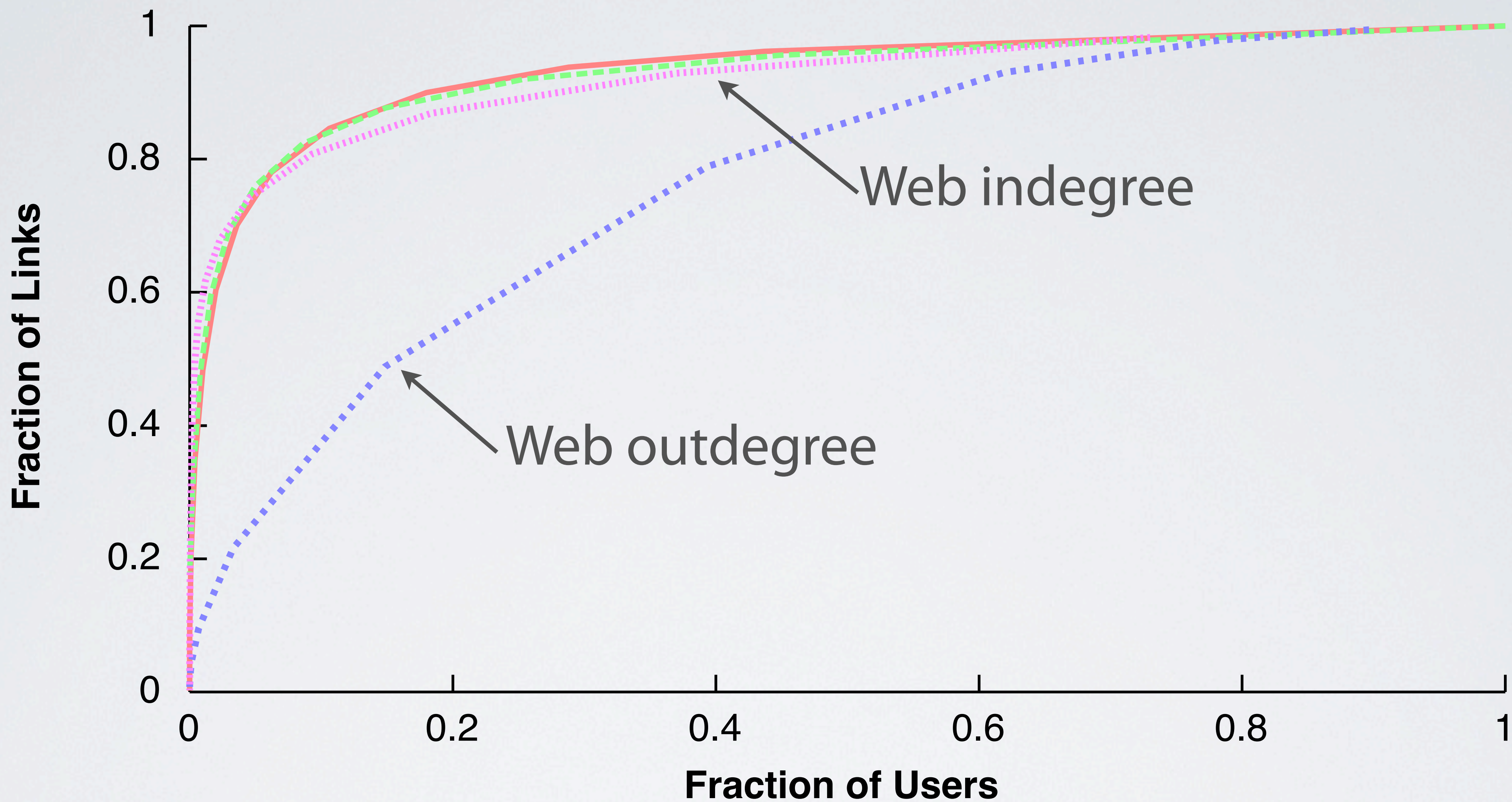
	Outdegree γ	Indegree γ
Flickr	1.74	1.78
LiveJournal	1.59	1.65
Orkut	1.50	1.50
YouTube	1.63	1.99
Web	2.67	2.09

Estimated coefficients with maximum likelihood testing

- Flickr, LiveJournal, YouTube have good K-S goodness-of-fit
- Orkut deviates due to partial crawl

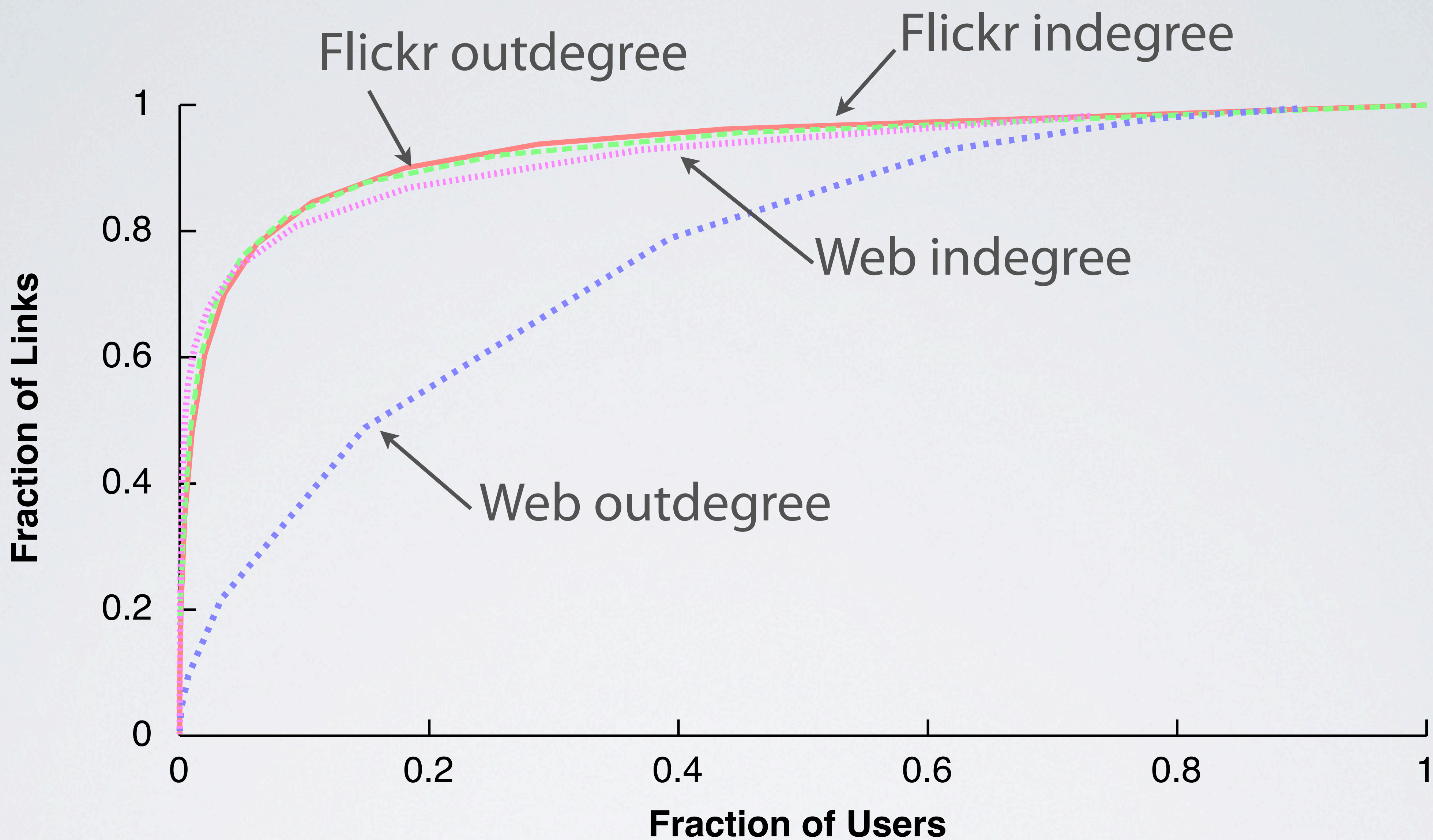
Similar coefficients imply a similar distribution of in/outdegree

How are the links distributed?



- Distribution of indegree and outdegree is similar, unlike Web
  - Underlying cause is link symmetry

How are the links distributed?



Distribution of indegree and outdegree is similar, unlike Web

Underlying cause is link symmetry

What fraction of links are symmetric?

Social networks show high level of link symmetry

- Even though links in most networks are directed

	Flickr	LiveJournal	Orkut	YouTube
Symmetric Links				

High symmetry increases network connectivity

- Reduces network diameter

What fraction of links are symmetric?

Social networks show high level of link symmetry

Even though links in most networks are directed

	Flickr	LiveJournal	Orkut	YouTube
Symmetric Links	62%	73%	100%	79%

High symmetry increases network connectivity

Reduces network diameter

Implications of high symmetry

High link symmetry implies indegree equals outdegree

- Users tend to receive as many links as the give

Unlike other complex networks, such as the Web

- Sites like cnn.com receive much links more than they give

Implications is that 'hubs' become 'authorities'

- May impact search algorithms (PageRank, HITS)

So far, observed networks are power-law with high symmetry

- Take a closer look next

Complex network structure

What is the high-level structure of online social networks?

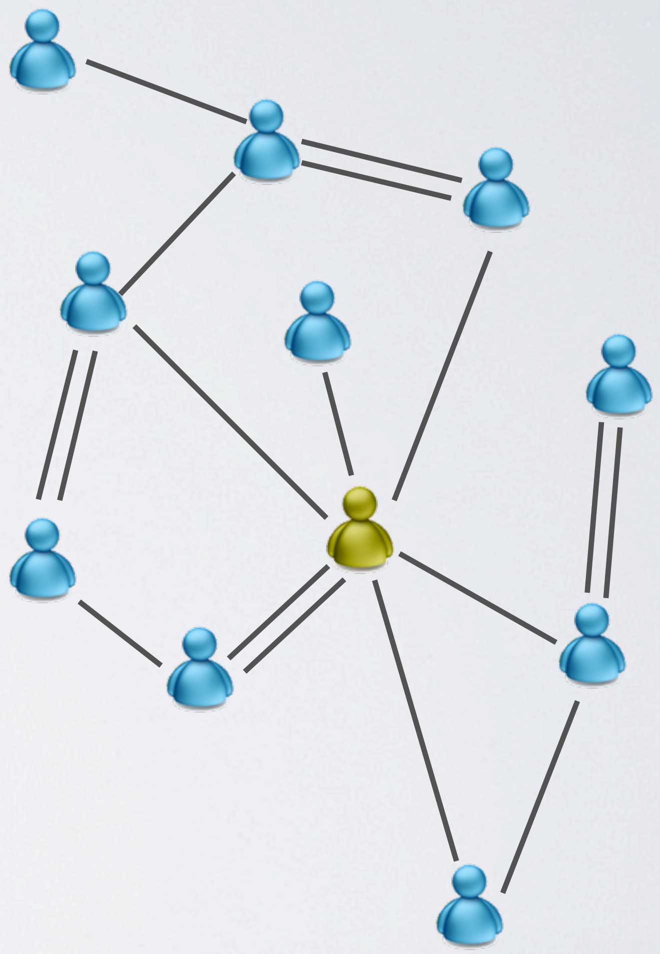
- A jellyfish, like the Internet? [Tauro et al, JCN 2001]
- A bowtie, like the Web? [Broder et al., WWW 2000]

In particular, is there a core of the network?

- Core is a (minimal) connected component
- Removing core disconnects remaining nodes

Approximate core detection by removing high-degree nodes

- When does network break apart?



Complex network structure

What is the high-level structure of online social networks?

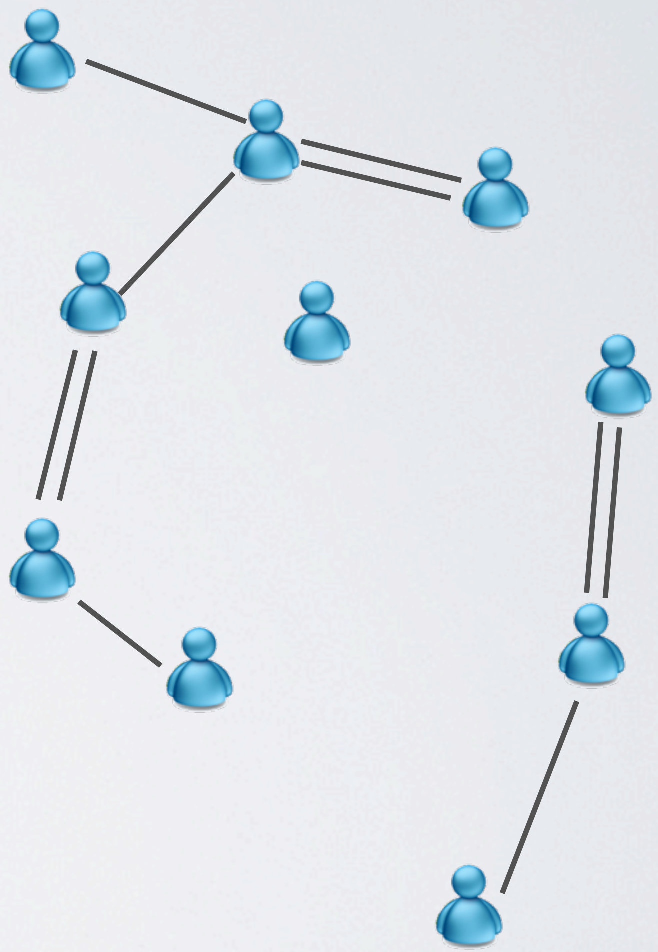
- A jellyfish, like the Internet? [Tauro et al, JCN 2001]
- A bowtie, like the Web? [Broder et al., WWW 2000]

In particular, is there a core of the network?

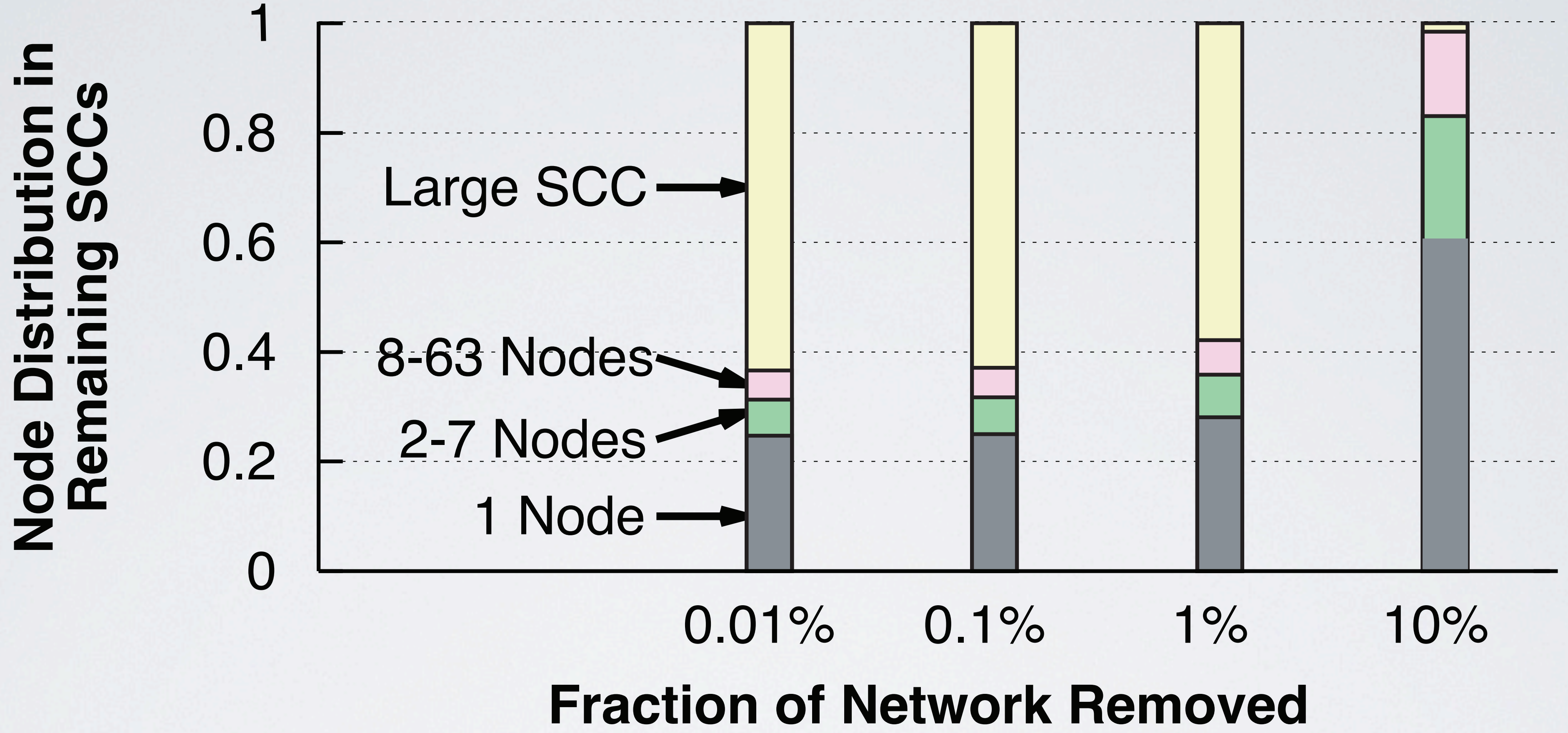
- Core is a (minimal) connected component
- Removing core disconnects remaining nodes

Approximate core detection by removing high-degree nodes

- When does network break apart?



Does a core exist?

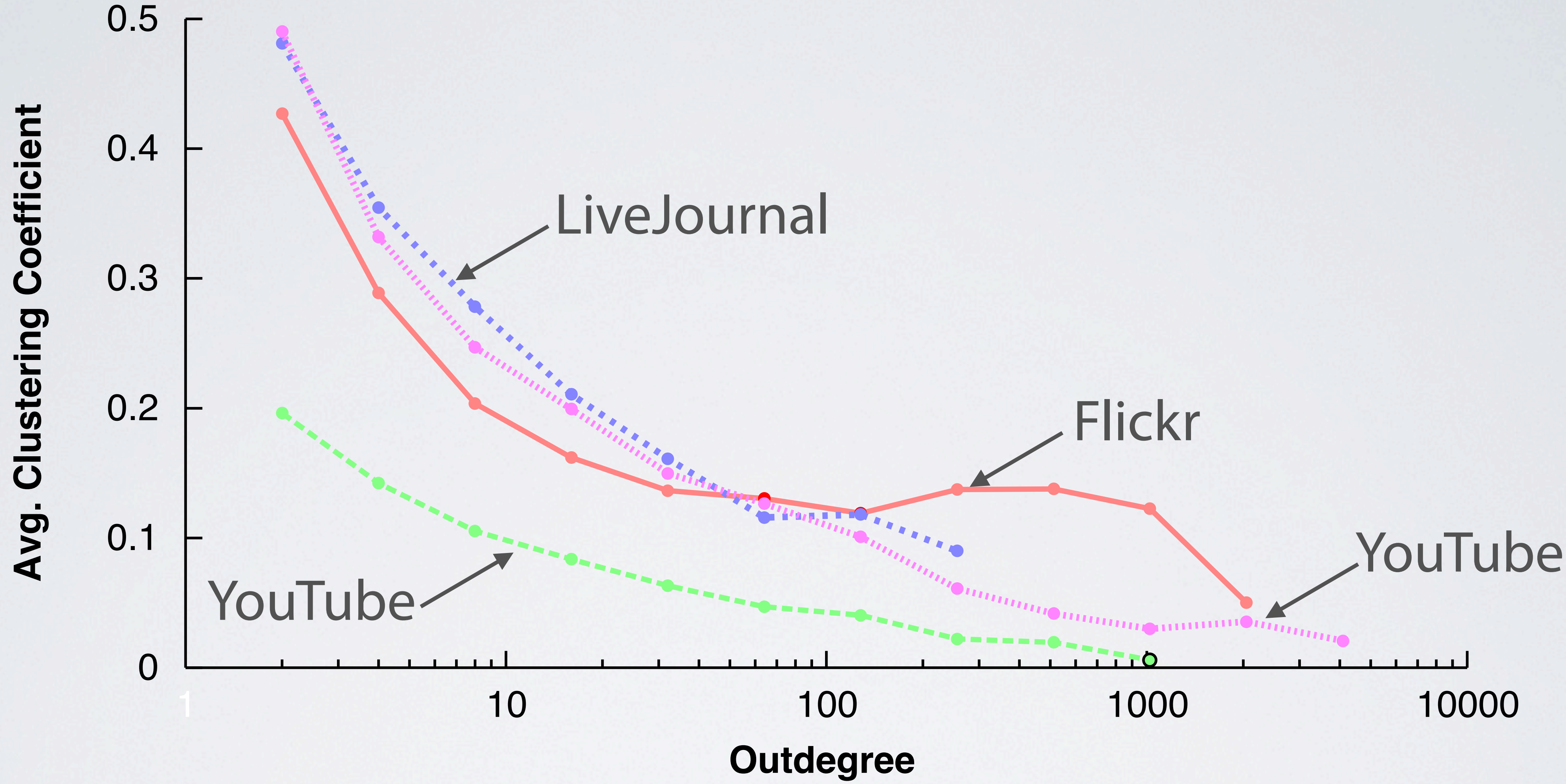


Yes, networks contain core consisting of 1-10% of nodes

- Removing core disconnects other nodes

What about remaining nodes (the fringe)?

How clustered is the fringe?



Low-degree users show high degree of clustering

Networks are small-world, may be scale-free

Implications

Disparate networks show similar overall structure

- Mechanisms and policies don't cause networks to look different

Network contains dense core of users

- Core necessary for connectivity of 90% of users
- Most short paths pass through core
- Could be used for quickly disseminating information

Fringe is highly clustered

- Users with few friends form mini-cliques
- Similar to previously observed offline behavior
- Could be leveraged for sharing information of local interest

User Interactions in Social Networks and their Implications

by Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. Puttaswamy, and Ben Y. Zhao

[Proceedings of EuroSys 2009]

User Interactions in Social Networks and their Implications

Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao
Computer Science Department, University of California at Santa Barbara
{bowlin, bboe, alessandra, krishnap, ravenben}@cs.ucsb.edu

Abstract

Social networks are popular platforms for interaction, communication and collaboration between friends. Researchers have recently proposed an emerging class of applications that leverage relationships from social networks to improve security and performance in applications such as email, web browsing and overlay routing. While these applications often cite social network connectivity statistics to support their designs, researchers in psychology and sociology have repeatedly cast doubt on the practice of inferring meaningful relationships from social network connections alone. This leads to the question: *Are social links valid indicators of real user interaction? If not, then how can we quantify these factors to form a more accurate model for evaluating socially-enhanced applications?* In this paper, we address this question through a detailed study of user interactions in the Facebook social network. We propose the use of *interaction graphs* to impart meaning to online social links by quantifying user interactions. We analyze interaction graphs derived from Facebook user traces and show that they exhibit significantly lower levels of the "small-world" properties shown in their social graph counterparts. This means that these graphs have fewer "supernodes" with extremely high degree, and overall network diameter increases significantly as a result. To quantify the impact of our observations, we use both types of graphs to validate two well-known social-based applications (RE [Garriss 2006] and SybilGuard [Yu 2006]). The results reveal new insights into both systems, and confirm our hypothesis that studies of social applications should use real indicators of user interactions in lieu of social graphs.

Categories and Subject Descriptors C.2.4 [*Distributed Systems*]: Distributed Applications

General Terms Measurement, Performance

EuroSys'09, April 1–3, 2009, Nuremberg, Germany.

1. Introduction

Social networks are popular infrastructures for communication, interaction, and information sharing on the Internet. Popular social networks such as MySpace and Facebook provide communication, storage and social applications for hundreds of millions of users. Users join, establish *social links* to friends, and leverage their social links to share content, organize events, and search for specific users or shared resources. These social networks provide platforms for organizing events, user to user communication, and are among the Internet's most popular destinations.

Recent work has seen the emergence of a class of socially-enhanced applications that leverage relationships from social networks to improve security and performance of network applications, including spam email mitigation [Garriss 2006], Internet search [Mislove 2006], and defense against Sybil attacks [Yu 2006]. In each case, meaningful, interactive relationships with friends are critical to improving trust and reliability in the system.

Unfortunately, these applications assume that all online social links denote a uniform level of real-world interpersonal association, an assumption disproven by social science. Specifically, social psychologists have long observed the prevalence of low-interaction social relationships such as Milgram's "Familiar Stranger" [Milgram 1977]. Recent research on social computing shows that users of social networks often use public display of connections to represent status and identity [Donath 2004], further supporting the hypothesis that social links often connect acquaintances with no level of mutual trust or shared interests.

This leads to the question: *Are social links valid indicators of real user interaction? If not, then what can we use to form a more accurate model for evaluating socially-enhanced applications?* In this paper, we address this question through a detailed study of user interaction events in Facebook, the most popular social network in the US with over 110 million active users. We download more than 10 million user profiles from Facebook, and examine records of user interactions to analyze interaction patterns across large user groups. Our results show that user interactions do in fact deviate significantly from social link patterns, in terms of factors such as time in the network, method of interaction, and types of users involved.

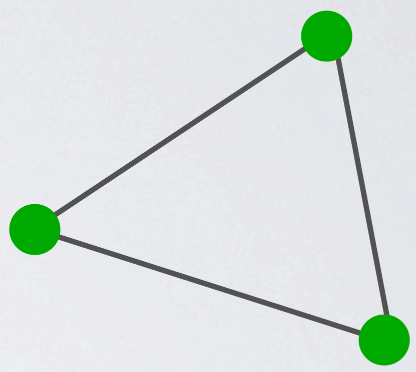
What do links mean?

Recall, social media defined by user interaction

- “Links” represent interacting users

But all links may not be equal, represent

- Best friends
- Acquantances
- Enemies
- Users who don’t know each other



Being “friends “ on social media sites requires very little effort

Question: What do social media links imply?

- Do these represent real “friends”?

This paper

Study interaction

- Wall posts, status comments, messages

Question: Are social links valid indicators of real user interaction?

- First large scale study of Facebook
- 10 million users (15% of total users)
- 24 million interactions

Use data to show highly skewed distribution of interactions

- <1% of people on Facebook talk to >50% of their friends

Collecting data

Crawling social networks is difficult

- Too large to crawl completely, must be sampled
- Privacy settings may prevent crawling

Facebook is divided into networks (regions, schools, companies)

- Regional networks not authenticated

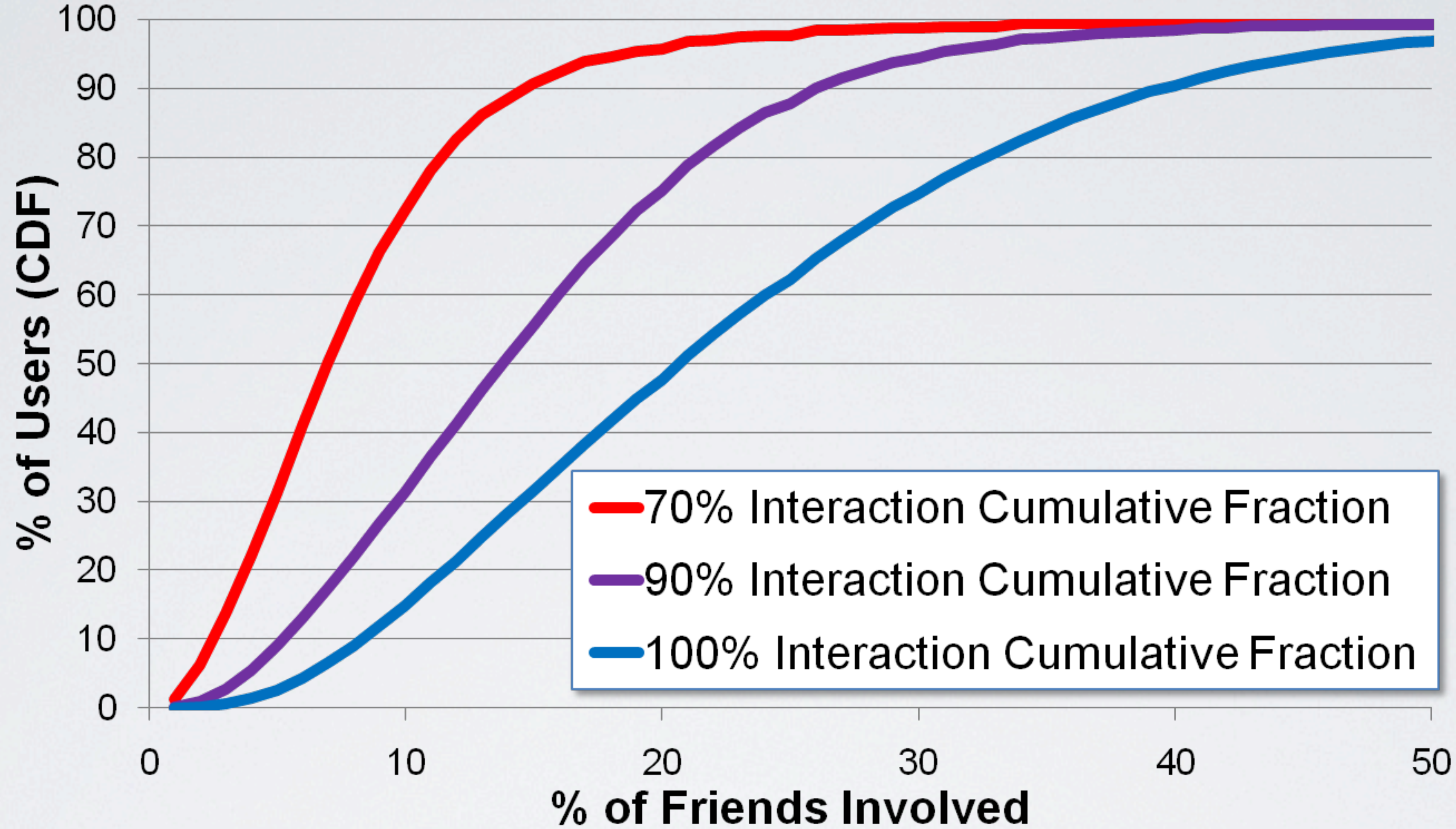
Crawled Facebook regional networks

- 22 largest networks: London, Australia, New York, ... (March – May 2008)
- Start with 50 random 'seed' users, perform BFS search

Data recorded for each user:

- Friends, wall posts, photo comments

How many social links have interaction?



- 50% of users interact with less than 20% of their friends
  - Many links never backed by interaction
- What if we only look at "interaction" links?

Interaction network

Not all social links are created equal

- Many (most?) social links are never "used"

What is the right way to model social networks?

- Take user interactivity into account

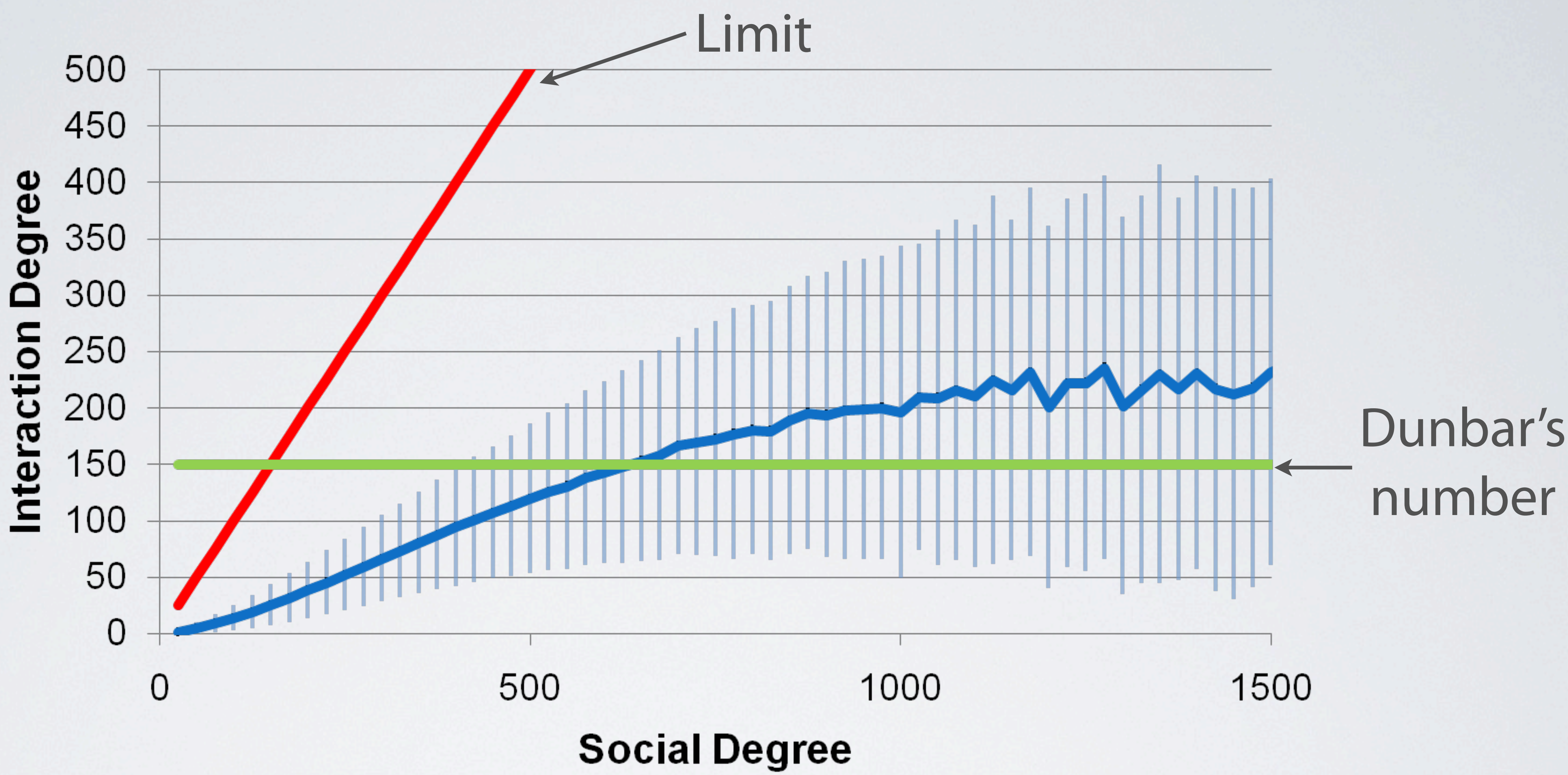
Interaction network: A social network parameterized by

- n : minimum number of interactions per link
- t : some window of time for interactions

For this study

- $n = 1$ and $t = \{2004 \text{ to the present}\}$

How many interaction links exist?



Interaction graph prunes unused edges

- Appears to be a limit in interaction degree
- Results agree with Dunbar's number

Summary

First large scale analysis of Facebook interaction

- Significant fraction of user population

Question: Are social links valid indicators of real user interaction?

- In general, no
- Formulate new model of social networks: Interaction network

Interaction networks have different characteristics than social networks

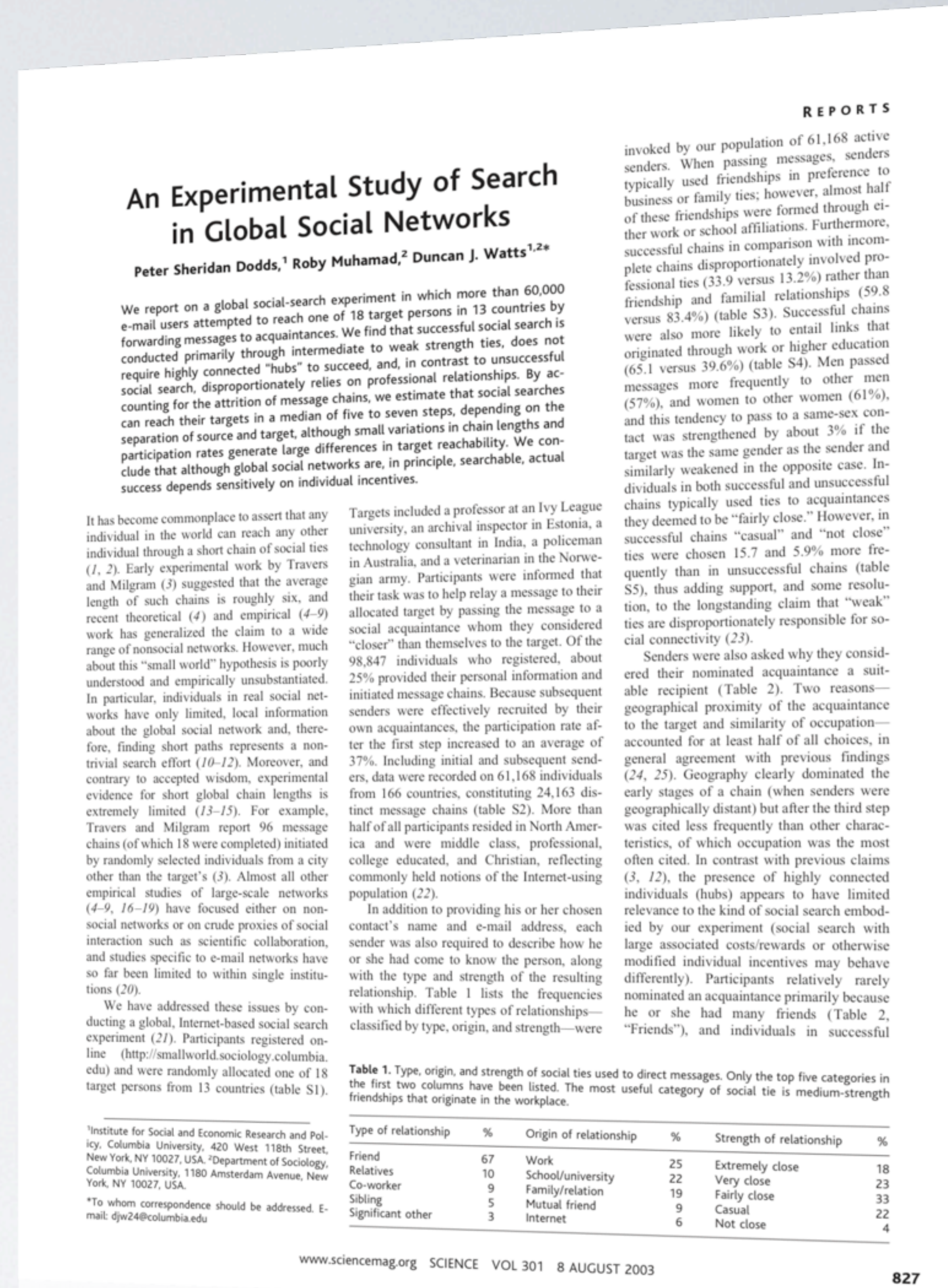
- More even distribution of links
- Bound on number of links per person

Maybe a better way to measure social networks?

An Experimental Study of Search in Global Social Networks

by Peter Sheridan Dodds, Roby Muhamad, and Duncan J. Watts

[Science, vol. 301 no. 5634. August 2003]



An Experimental Study of Search in Global Social Networks

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

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

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

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

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

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

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

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

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

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

Questions unanswered in Milgram's study

How did user pick next hop?

- Limited, local information
- No global view exists
- How did users find short paths?
- What made networks navigable?

Are results only for US?

- Milgram's study had only one destination
- Generalizable to different sources/destinations?
- What about languages, cultures, etc?

Experimental design

- Internet-based social search experiment
  - Replicate Milgram's study using social media
  - Much cheaper to do today
- Participants registered online and were allocated target
  - There were 18 target persons from 13 countries
- Similar instructions to Milgram
  - Relay a message to their target
  - Pass message only acquaintance they knew personally
  - Considered "closer" than themselves to the target

Experimental design

Internet-based social search experiment

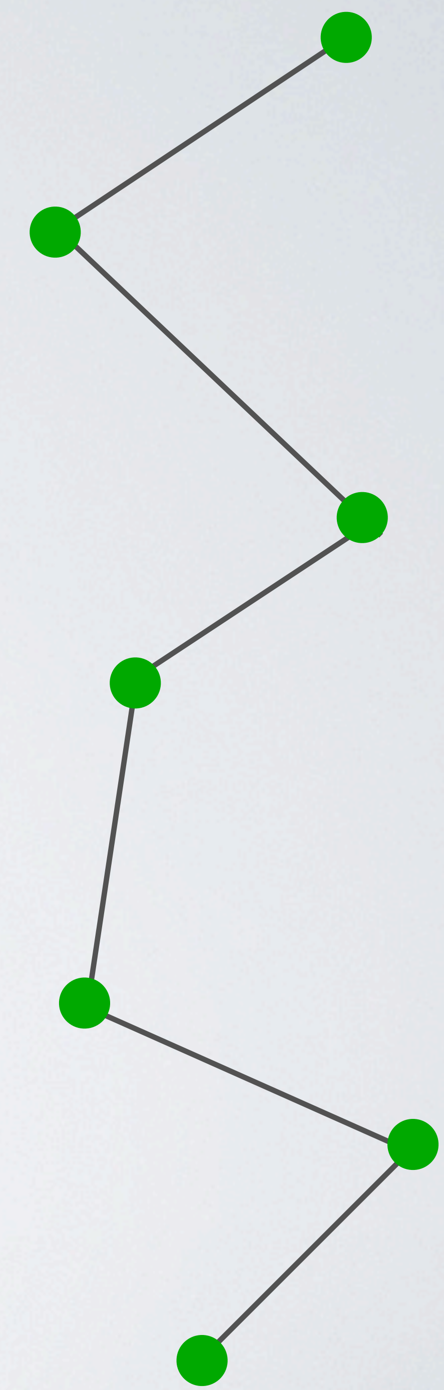
- Replicate Milgram's study using social media
- Much cheaper to do today

Participants registered online and were allocated target

- There were 18 target persons from 13 countries

Similar instructions to Milgram

- Relay a message to their target
- Pass message only acquaintance they knew personally
- Considered "closer" than themselves to the target



Data collected

- 98,847 individuals registered
  - 25% initiated message chains
  - Participation rate after the first step was 37%
- 24,163 distinct message chains
  - Included total of 61,168 individuals from 166 countries
  - Two orders of magnitude more than Milgram
- Information collected about the links
  - How user had come to know the other person
  - Type and strength of the relationship
  - Why they considered their nominated acquaintance a suitable recipient

Who did users pick?

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

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

- Friendships used in preference to business or family ties
  - Half formed through either work or school affiliations
- In successful chains, non-close ties chosen more
  - “Weak” ties are responsible for social connectivity
  - Bridge communities

How did this change?

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

- Users tend to use geography early in the path
  - Try and get message to the right region
- Then, switch to other attributes
  - Work, education, ...

Summary

- Replicated Milgram's study using social media
  - Can shed light on unanswered questions
- Do results generalize?
  - Found median chain length of 7, agrees well
- How did users route?
  - Geography dominated early
  - Work and education dominated later
- Provides insight into structure of navigable social networks

Part III

Leveraging social media

- or -

What is this all good for?

Three papers on leveraging social media

Predicting the Future With Social Media

You Are Who You Know: Inferring User Profiles in Online Social Networks

Meme-tracking and the Dynamics of the News Cycle

Cover three topics

- Tracking information flow
- Applying social media to real-world problems
- Privacy implications of social media

Meme-tracking and the Dynamics of the News Cycle

by Jure Leskovec, Lars Backstrom, and 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

KDD'09, June 28–July 1, 2009, Paris, France.

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

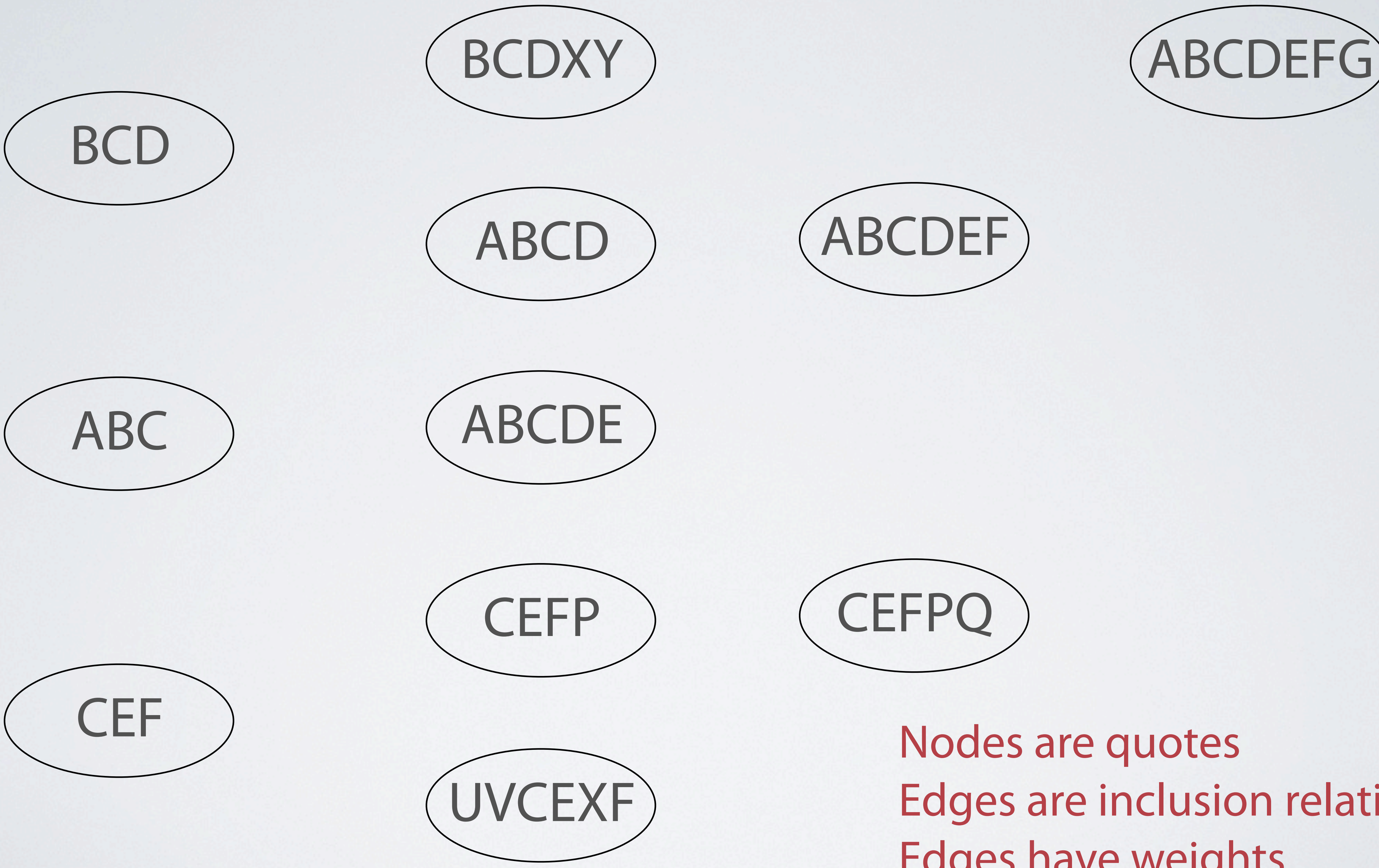
spinn3r

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..."

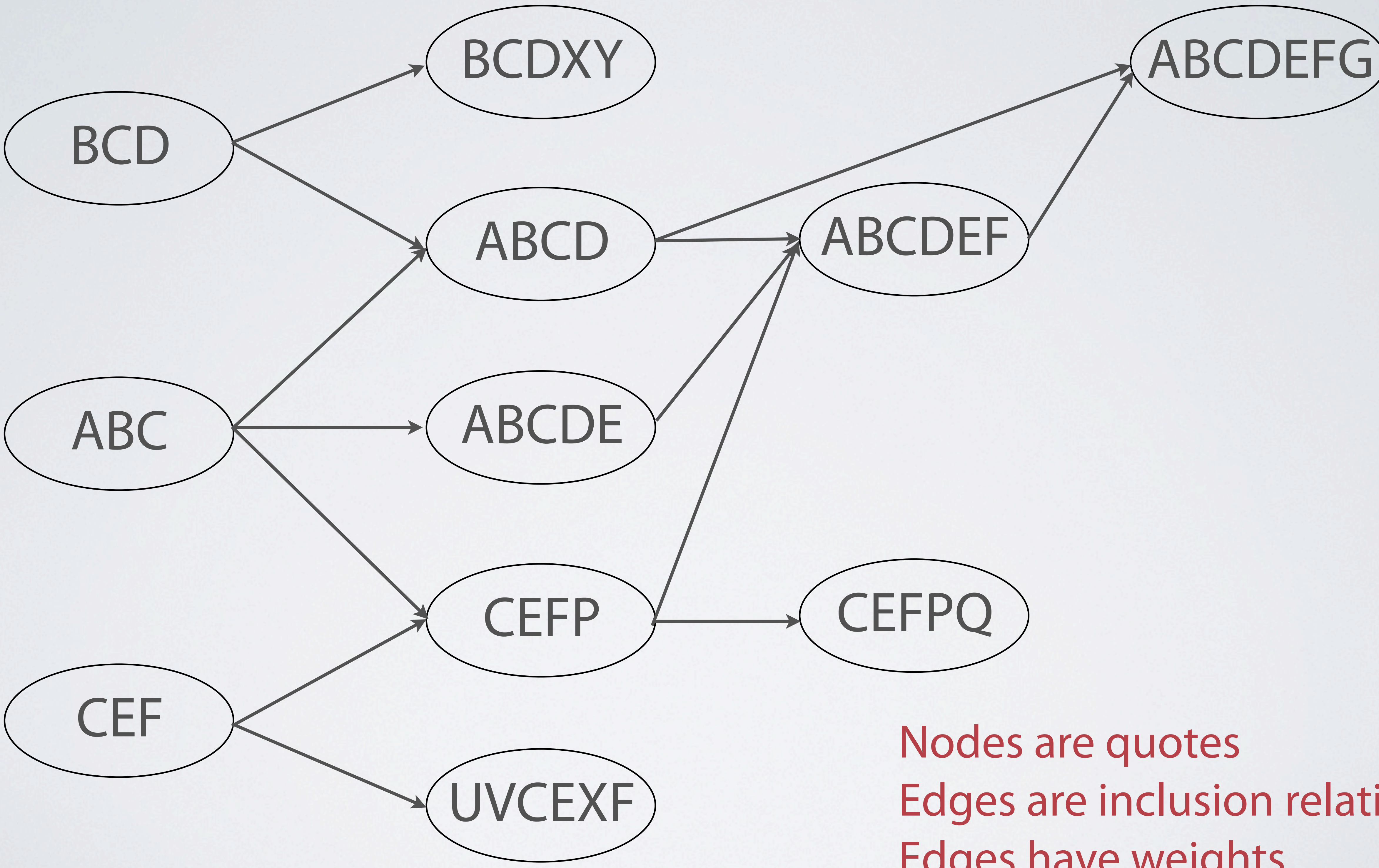
How to determine which quotes are the same?

Clustering quotes



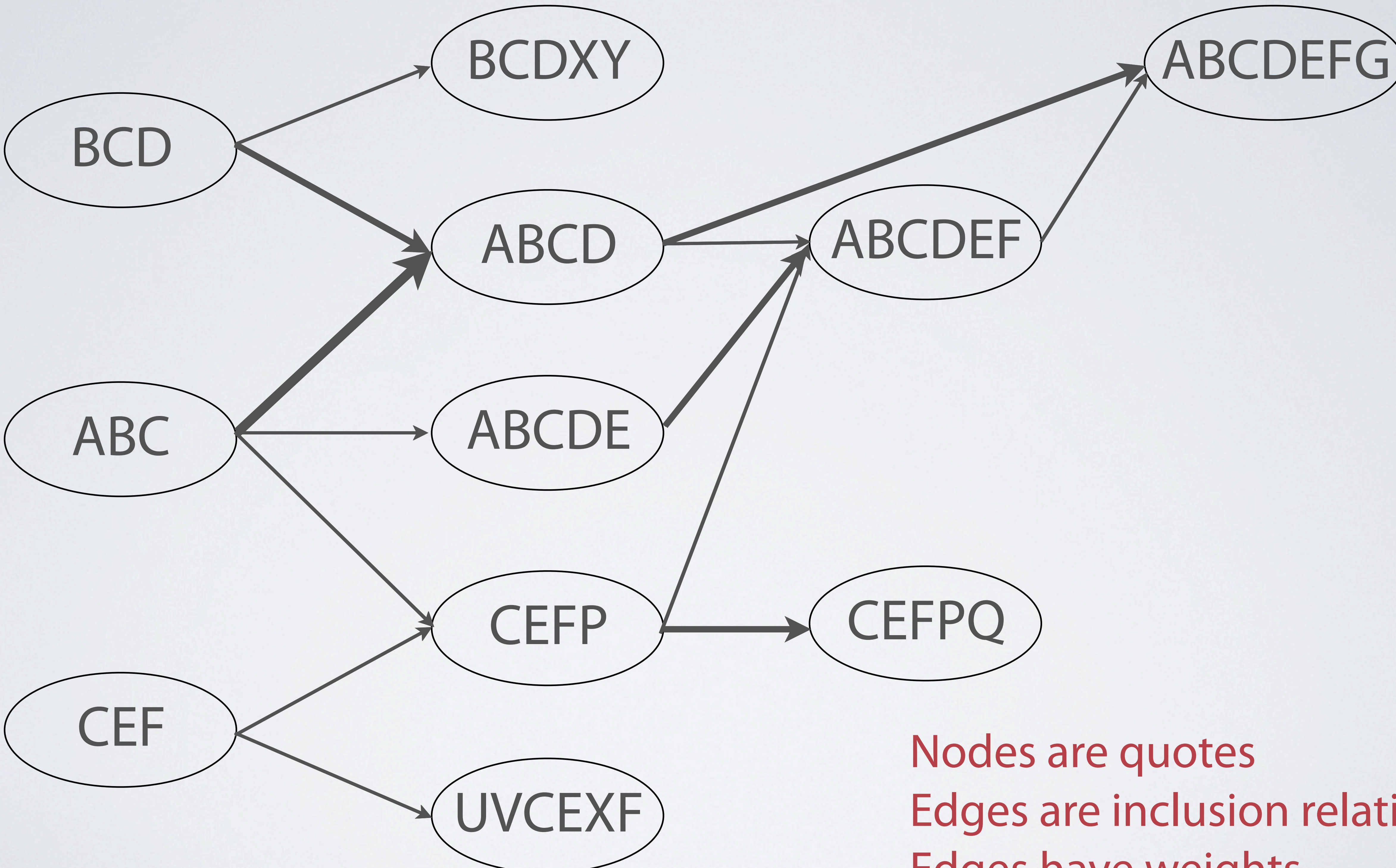
Nodes are quotes
Edges are inclusion relations
Edges have weights

Clustering quotes



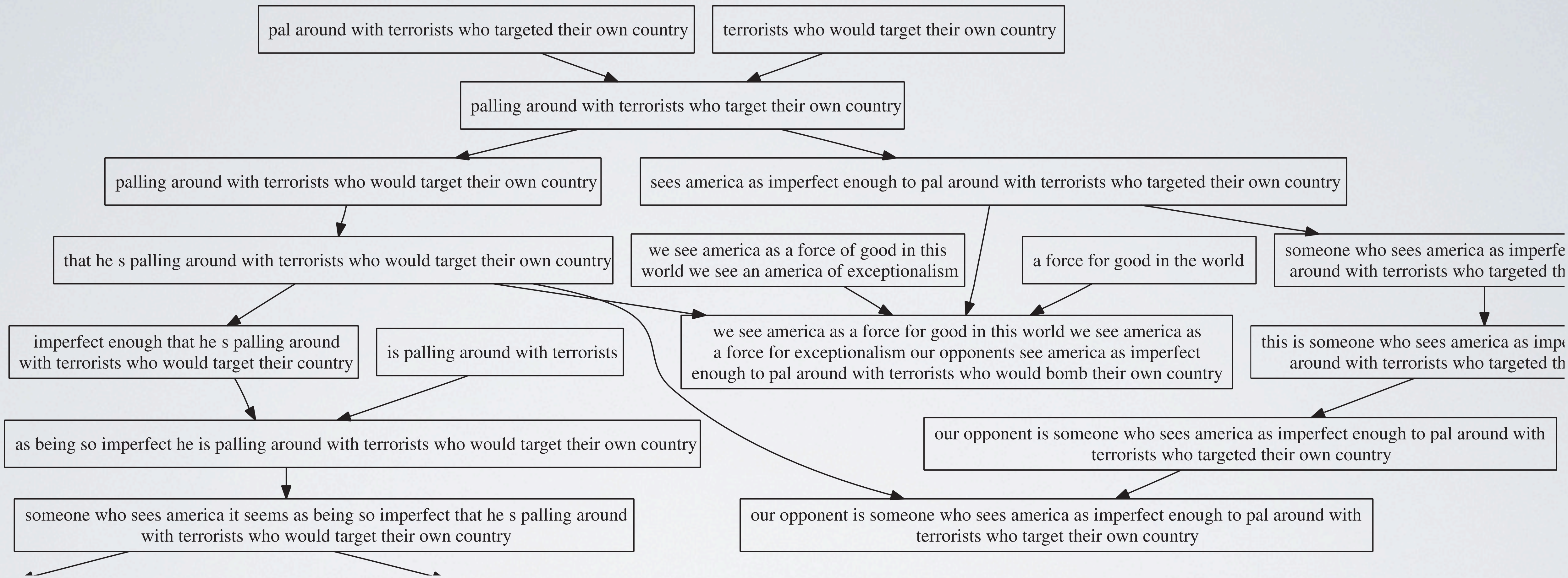
Nodes are quotes
Edges are inclusion relations
Edges have weights

Clustering quotes



Nodes are quotes
Edges are inclusion relations
Edges have weights

Example of cluster

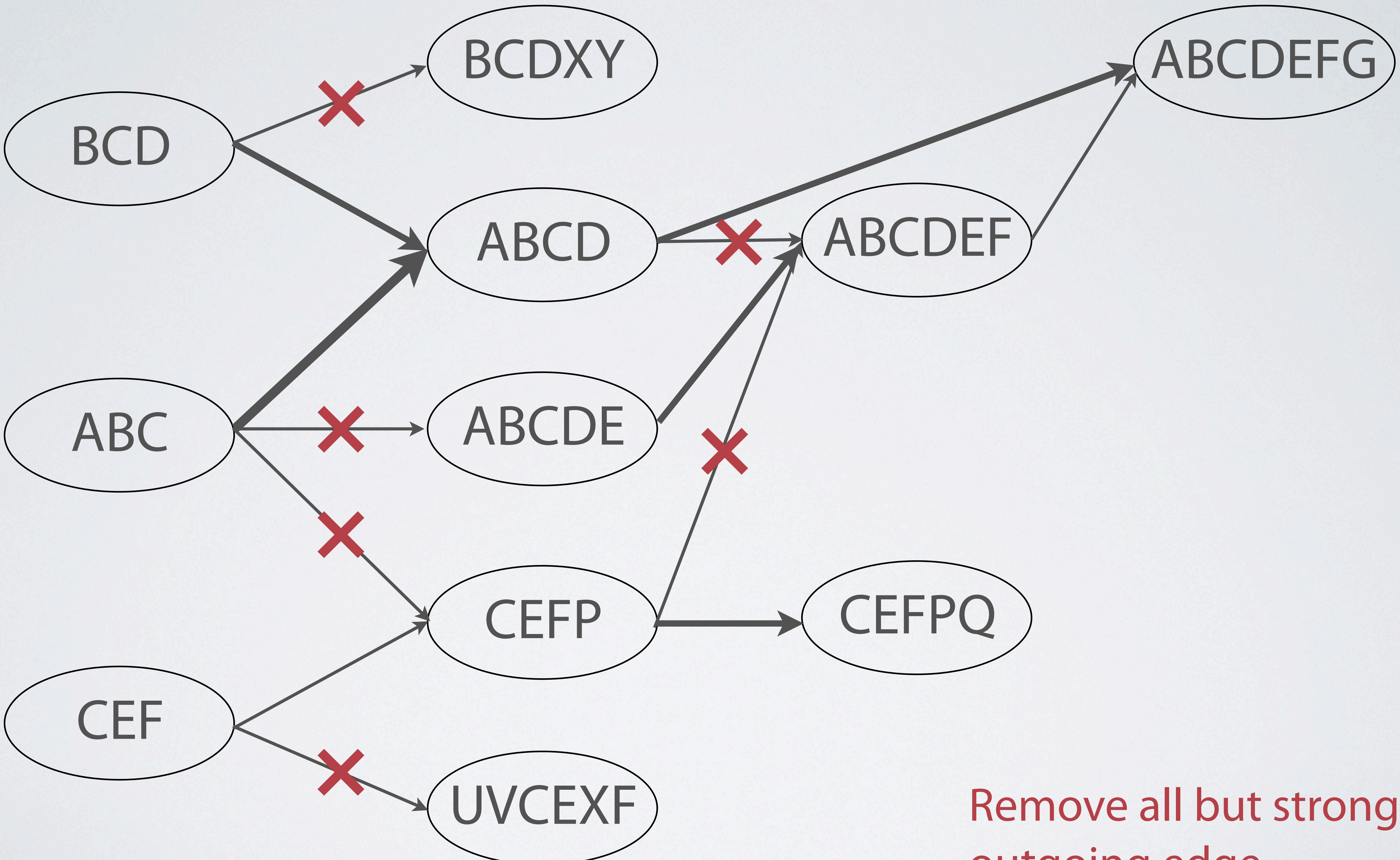


All based on Sarah Palin's terrorists quote:

"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."

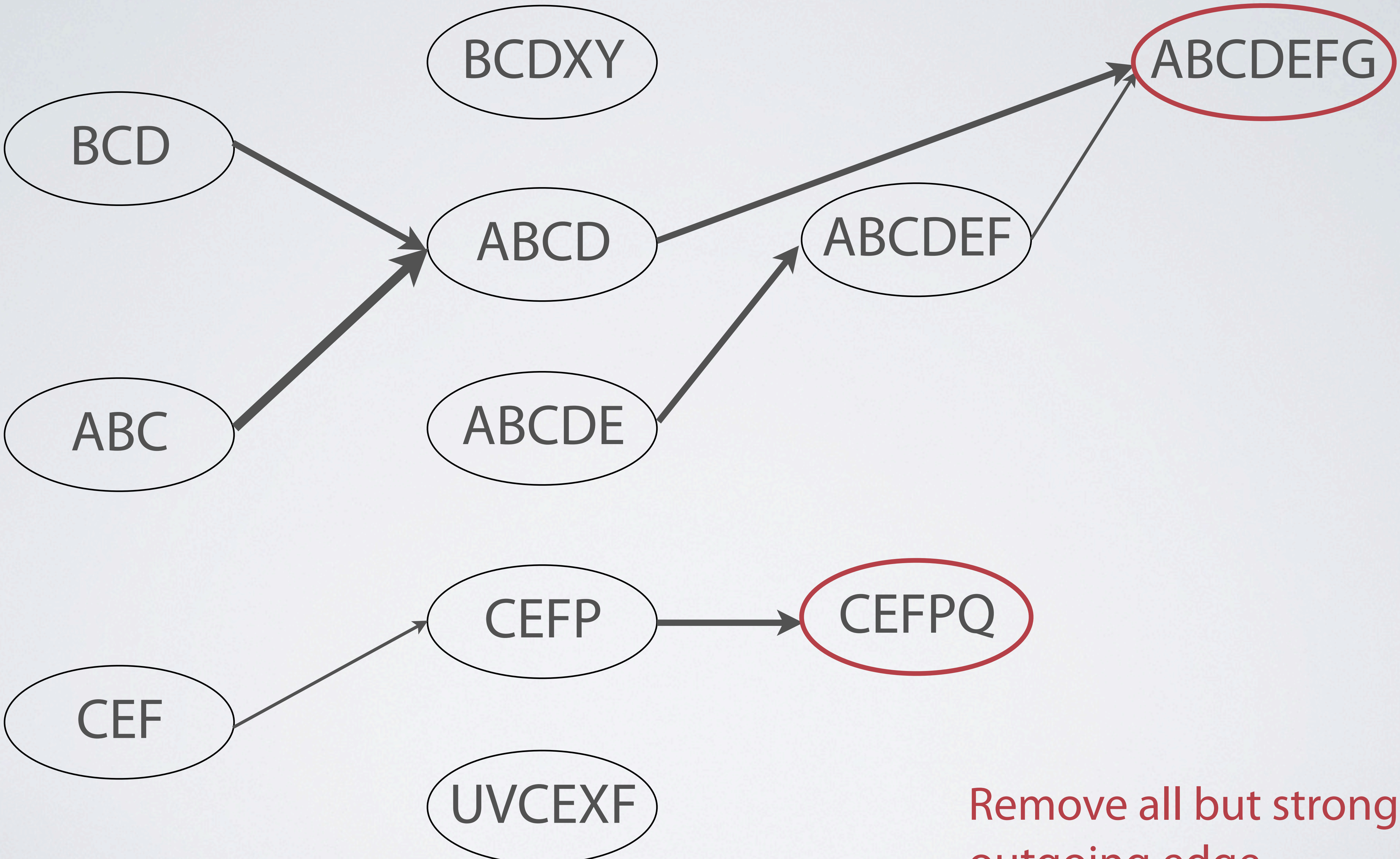
How to reduce to a single meme?

Solution: Create a DAG



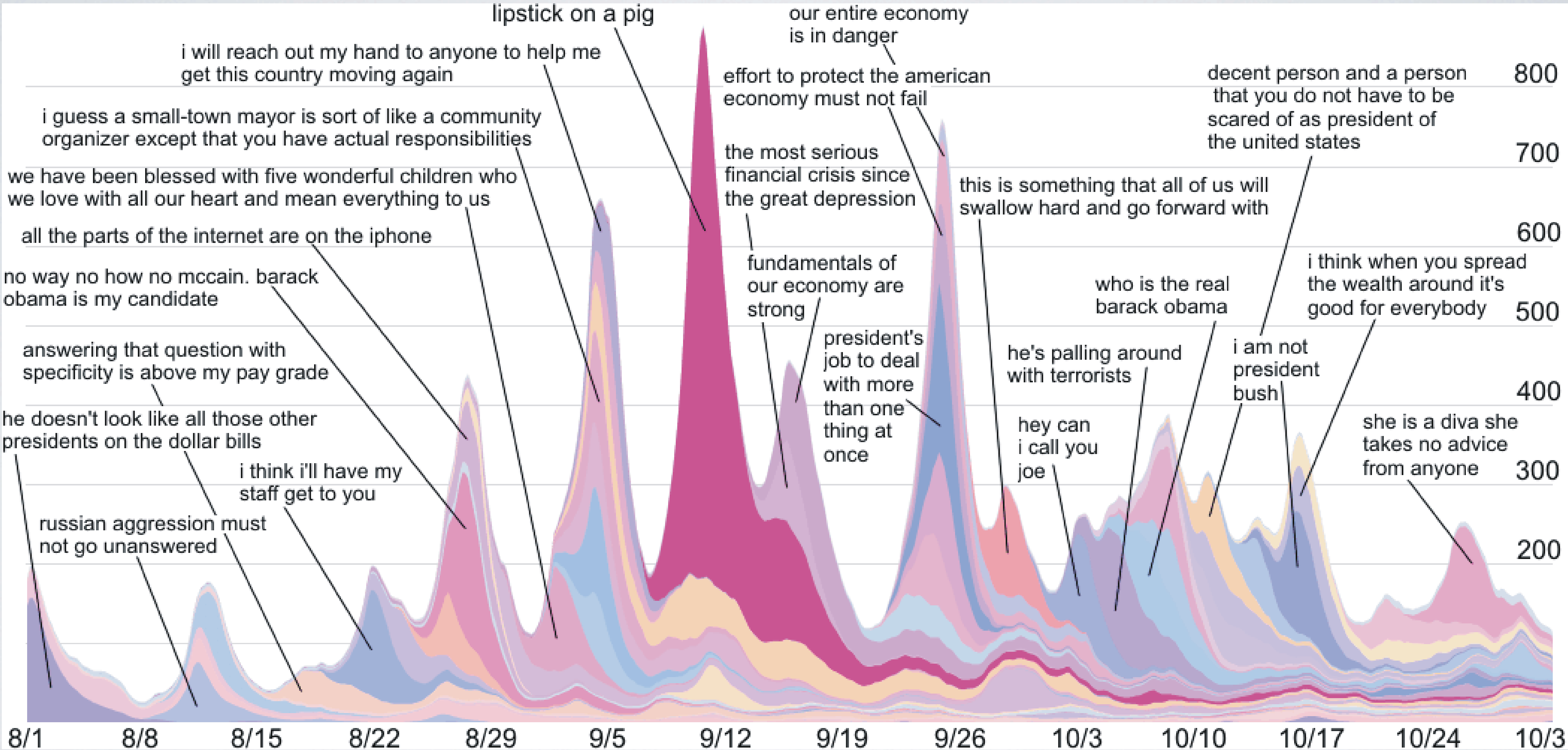
Remove all but strongest outgoing edge

Solution: Create a DAG



Remove all but strongest outgoing edge

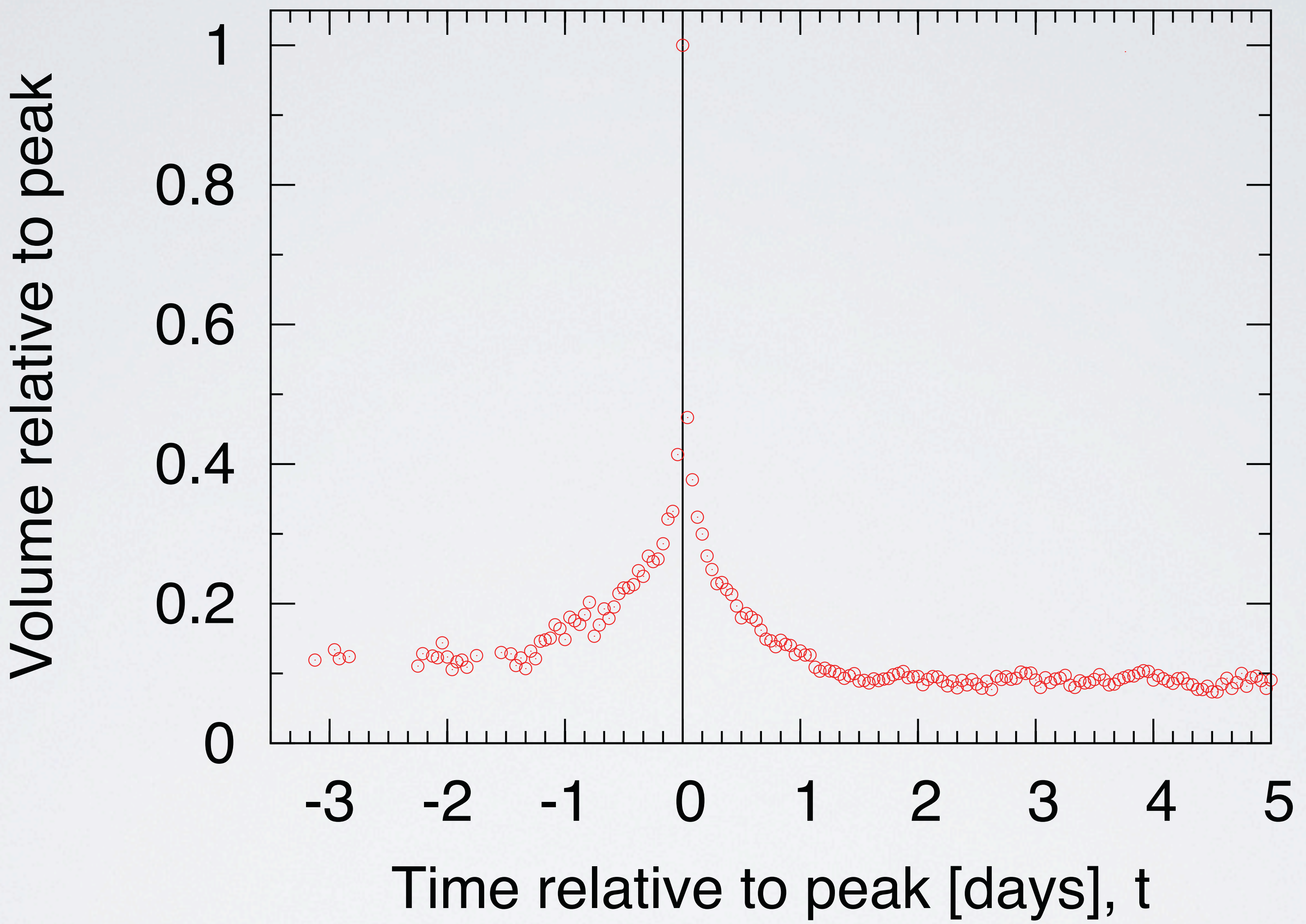
Resulting memes



Spikes show nature of 24-hour news cycle

Memes quickly enter and leave collective conscience

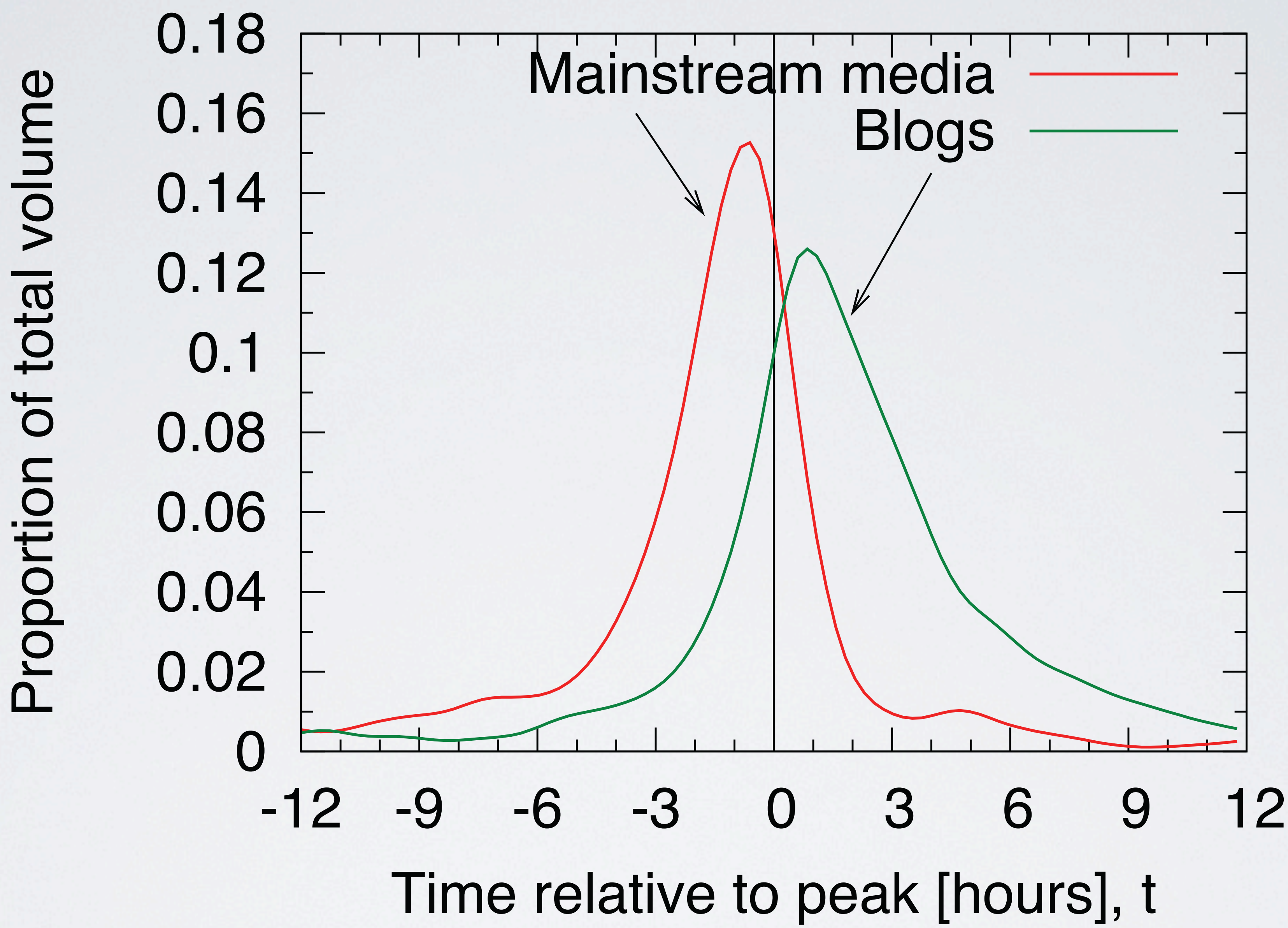
Tracking memes



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

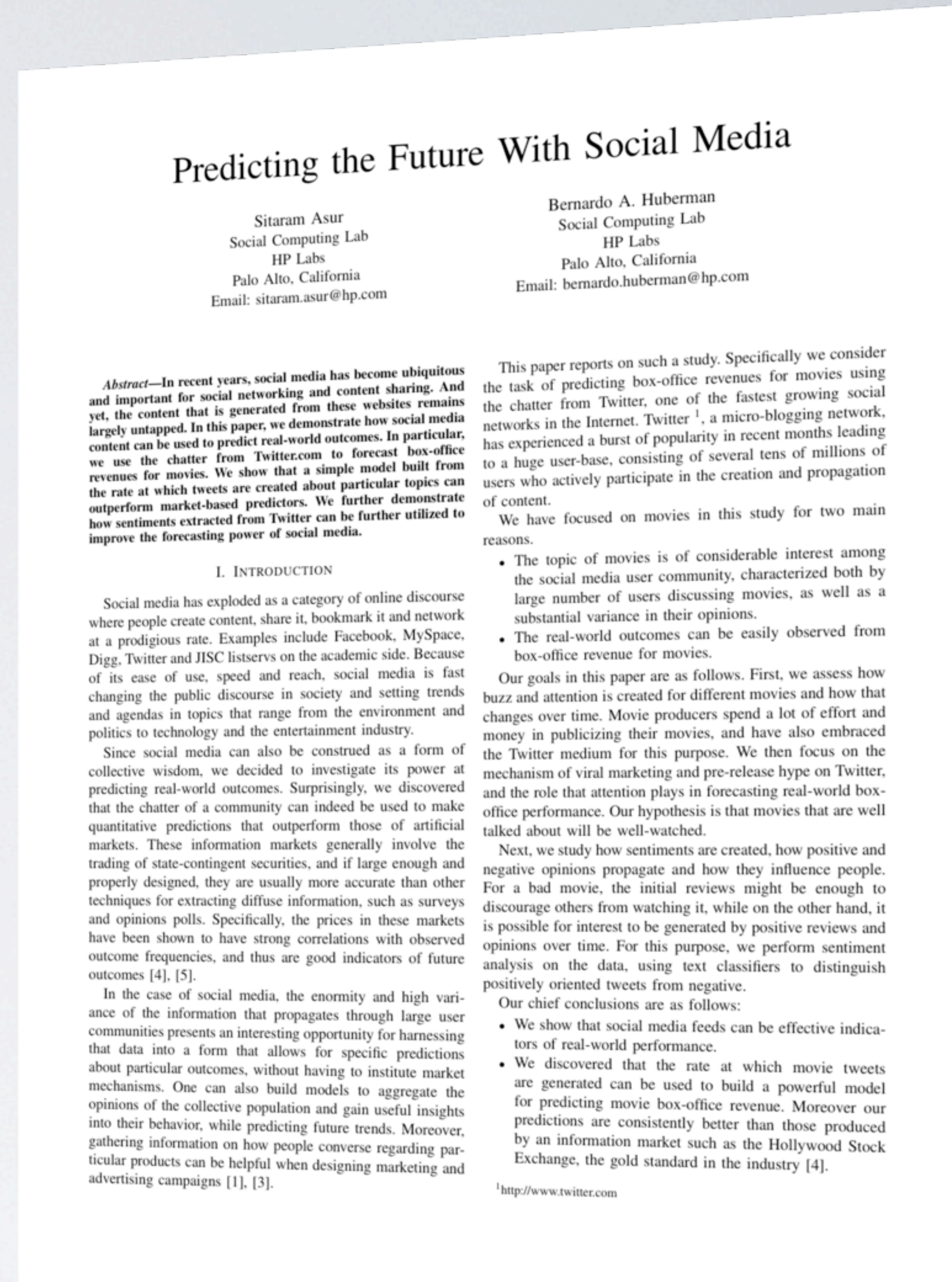
Found interesting interaction between media and blogs

- 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?

Armored
Avatar
The Blind Side
The Book of Eli
Daybreakers
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
Legion
Twilight : New Moon
Pirate Radio
Princess And The Frog
Sherlock Holmes
Spy Next Door
The Crazies
Tooth Fairy
Transylmania
When In Rome
Youth In Revolt

Focus on mentions of 24 movies on twitter

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

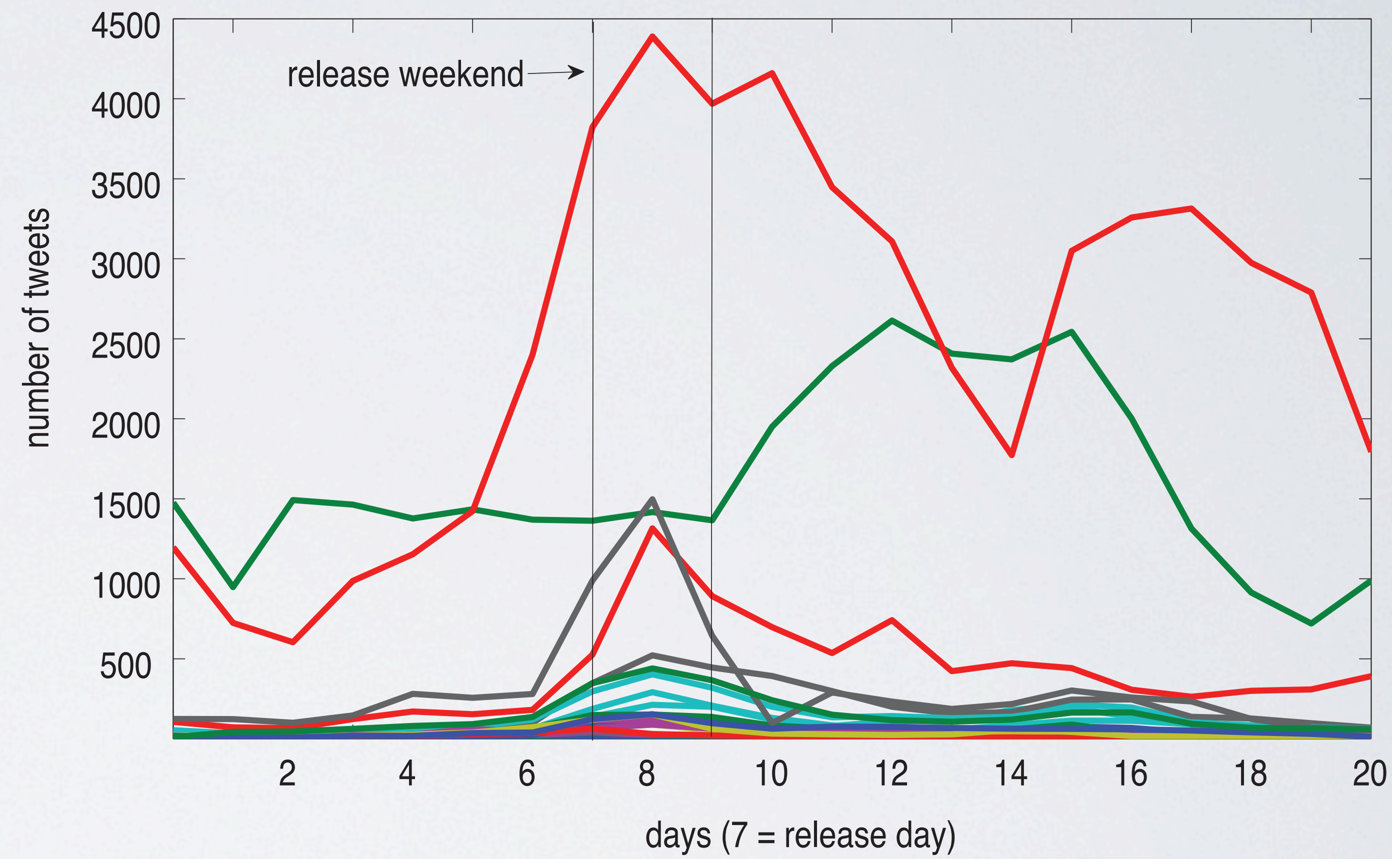
Busiest time is around release

- Promotions, advertising, ...

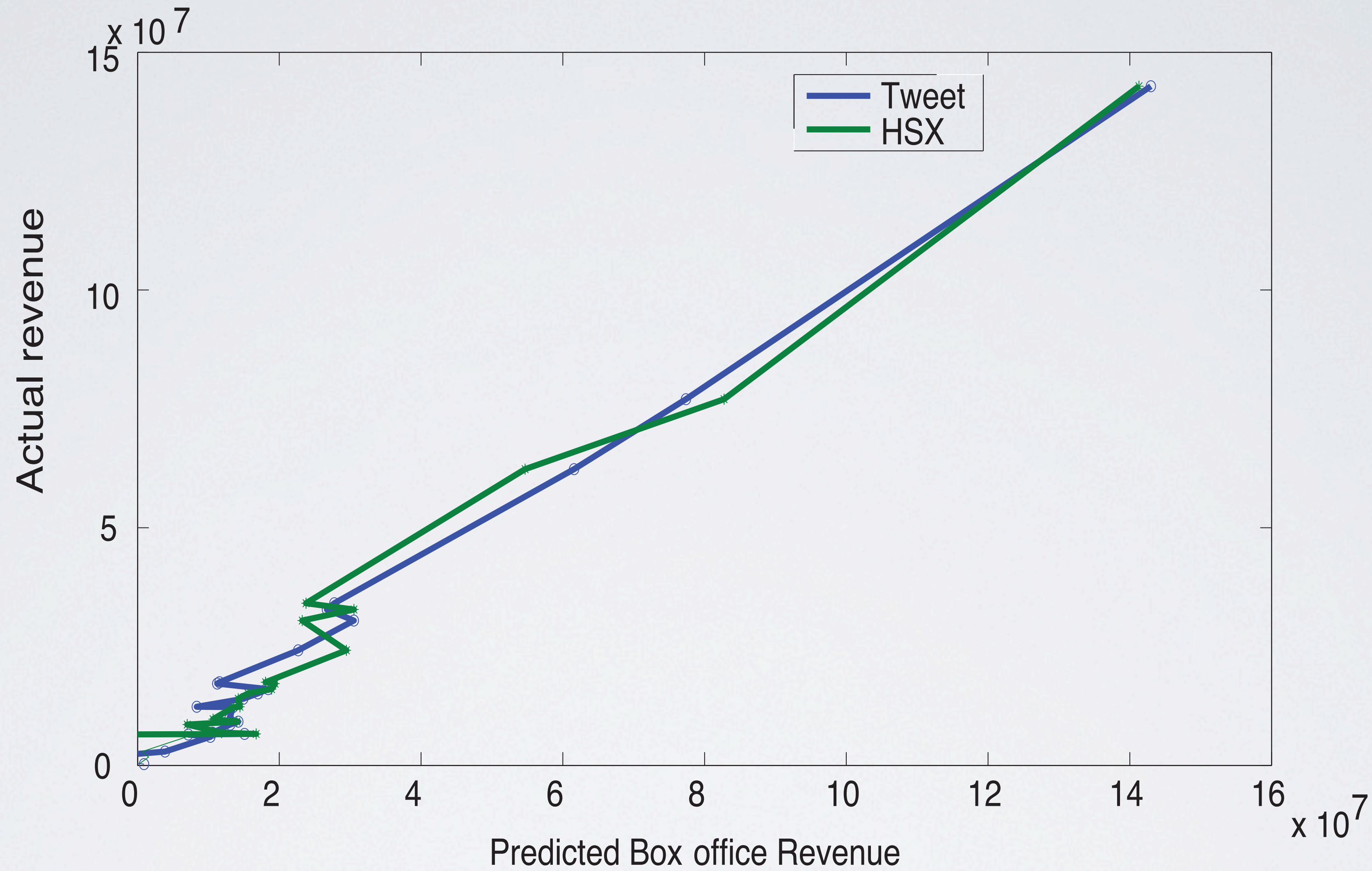
Opening weekend makes most money

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?

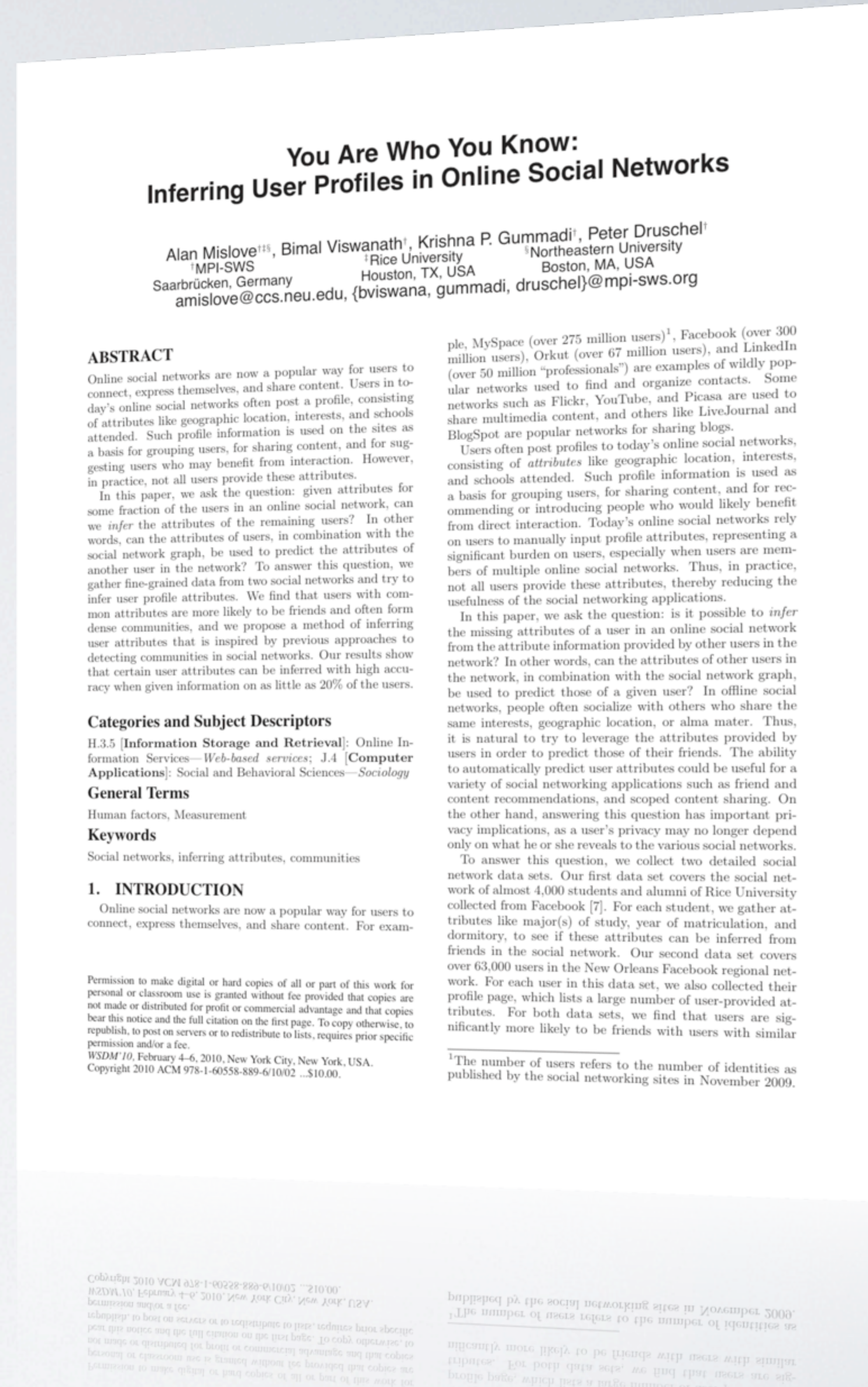
But unclear causality

- Do movie studios only promote movies they expect to be a hit?
- What about duds?

You Are Who You Know: Inferring User Profiles in Online Social Networks

by Alan Mislove, Bimal Viswanath, Krishna P. Gummadi, and Peter Druschel

[Proceedings of WSDM 2010]



You Are Who You Know: Inferring User Profiles in Online Social Networks

Alan Mislove[†‡§], Bimal Viswanath[†], Krishna P. Gummadi[†], Peter Druschel[†]
[†]MPI-SWS, Saarbrücken, Germany
[‡]Rice University, Houston, TX, USA
[§]Northeastern University, Boston, MA, USA
amislove@ccs.neu.edu, {bviswana, gummadi, druschel}@mpi-sws.org

ABSTRACT

Online social networks are now a popular way for users to connect, express themselves, and share content. Users in today's online social networks often post a profile, consisting of attributes like geographic location, interests, and schools attended. Such profile information is used on the sites as a basis for grouping users, for sharing content, and for suggesting users who may benefit from interaction. However, in practice, not all users provide these attributes.

In this paper, we ask the question: given attributes for some fraction of the users in an online social network, can we *infer* the attributes of the remaining users? In other words, can the attributes of users, in combination with the social network graph, be used to predict the attributes of another user in the network? To answer this question, we gather fine-grained data from two social networks and try to infer user profile attributes. We find that users with common attributes are more likely to be friends and often form dense communities, and we propose a method of inferring user attributes that is inspired by previous approaches to detecting communities in social networks. Our results show that certain user attributes can be inferred with high accuracy when given information on as little as 20% of the users.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology*

General Terms

Human factors, Measurement

Keywords

Social networks, inferring attributes, communities

1. INTRODUCTION

Online social networks are now a popular way for users to connect, express themselves, and share content. For example, MySpace (over 275 million users)[1], Facebook (over 300 million users), Orkut (over 67 million users), and LinkedIn (over 50 million "professionals") are examples of wildly popular networks used to find and organize contacts. Some networks such as Flickr, YouTube, and Picasa are used to share multimedia content, and others like LiveJournal and BlogSpot are popular networks for sharing blogs.

Users often post profiles to today's online social networks, consisting of *attributes* like geographic location, interests, and schools attended. Such profile information is used as a basis for grouping users, for sharing content, and for recommending or introducing people who would likely benefit from direct interaction. Today's online social networks rely on users to manually input profile attributes, representing a significant burden on users, especially when users are members of multiple online social networks. Thus, in practice, not all users provide these attributes, thereby reducing the usefulness of the social networking applications.

In this paper, we ask the question: is it possible to *infer* the missing attributes of a user in an online social network from the attribute information provided by other users in the network? In other words, can the attributes of other users in the network, in combination with the social network graph, be used to predict those of a given user? In offline social networks, people often socialize with others who share the same interests, geographic location, or alma mater. Thus, it is natural to try to leverage the attributes provided by users in order to predict those of their friends. The ability to automatically predict user attributes could be useful for a variety of social networking applications such as friend and content recommendations, and scoped content sharing. On the other hand, answering this question has important privacy implications, as a user's privacy may no longer depend only on what he or she reveals to the various social networks.

To answer this question, we collect two detailed social network data sets. Our first data set covers the social network of almost 4,000 students and alumni of Rice University collected from Facebook [7]. For each student, we gather attributes like major(s) of study, year of matriculation, and dormitory, to see if these attributes can be inferred from friends in the social network. Our second data set covers over 63,000 users in the New Orleans Facebook regional network. For each user in this data set, we also collected their profile page, which lists a large number of user-provided attributes. For both data sets, we find that users are significantly more likely to be friends with users with similar

[1]The number of users refers to the number of identities as published by the social networking sites in November 2009.

WSDM'10, February 4–6, 2010, New York City, New York, USA.

Social media and privacy

Users upload information to social media sites

- Profile information
- Status updates
- Photos, videos

Privacy model for data

- Choose what to reveal
- And what to keep private

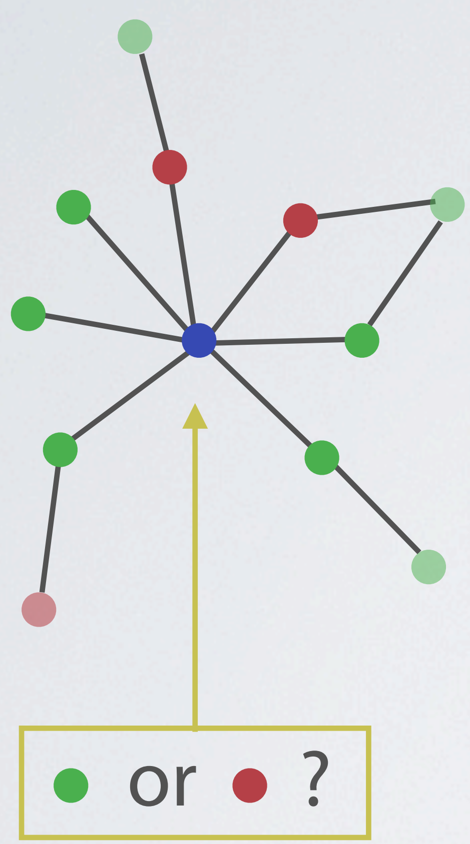
When reasoning about privacy

- Don't often consider implicit data
- What structure of the network reveals about us

facebook

+

What is implicit data?



Example: MIT's Project Gaydar

- Predict sexual orientation based on friends

Exploiting homophily

- People associate with others like them

What about other attributes?

Using friends-of-friends?

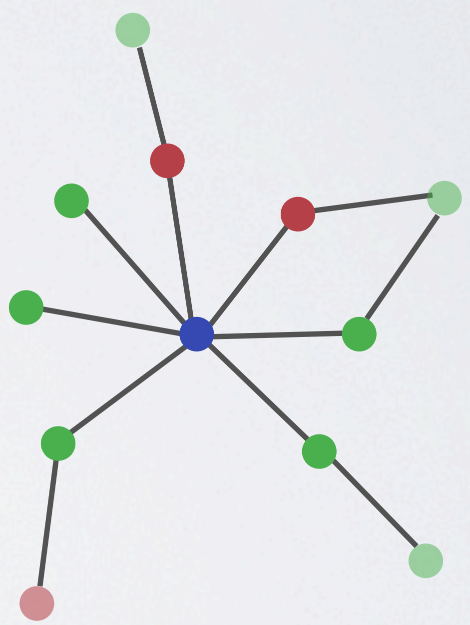
This paper

Explore how much implicit data exists on online social networks?

- Or, how much information can be inferred?
- How much data is needed to be able to infer?

Idea: Use communities

- Project Gaydar used 1-hop friends
- Using >1 hop friends is challenging
  - Exponential growth in size
  - Unclear relationship to source



Look for groupings of users

- Called communities
- Potentially share attributes

Social network data

Crawled two Facebook networks

- Rice University (university)
- New Orleans (regional)

Rice: authoritative attributes

- Queried student directory
- College (dormitory), major(s), year

New Orleans: extracted attributes from profile

- High school, favorite movies, birthday
- Non-authoritative, incomplete, freeform text

Do attributes define communities?

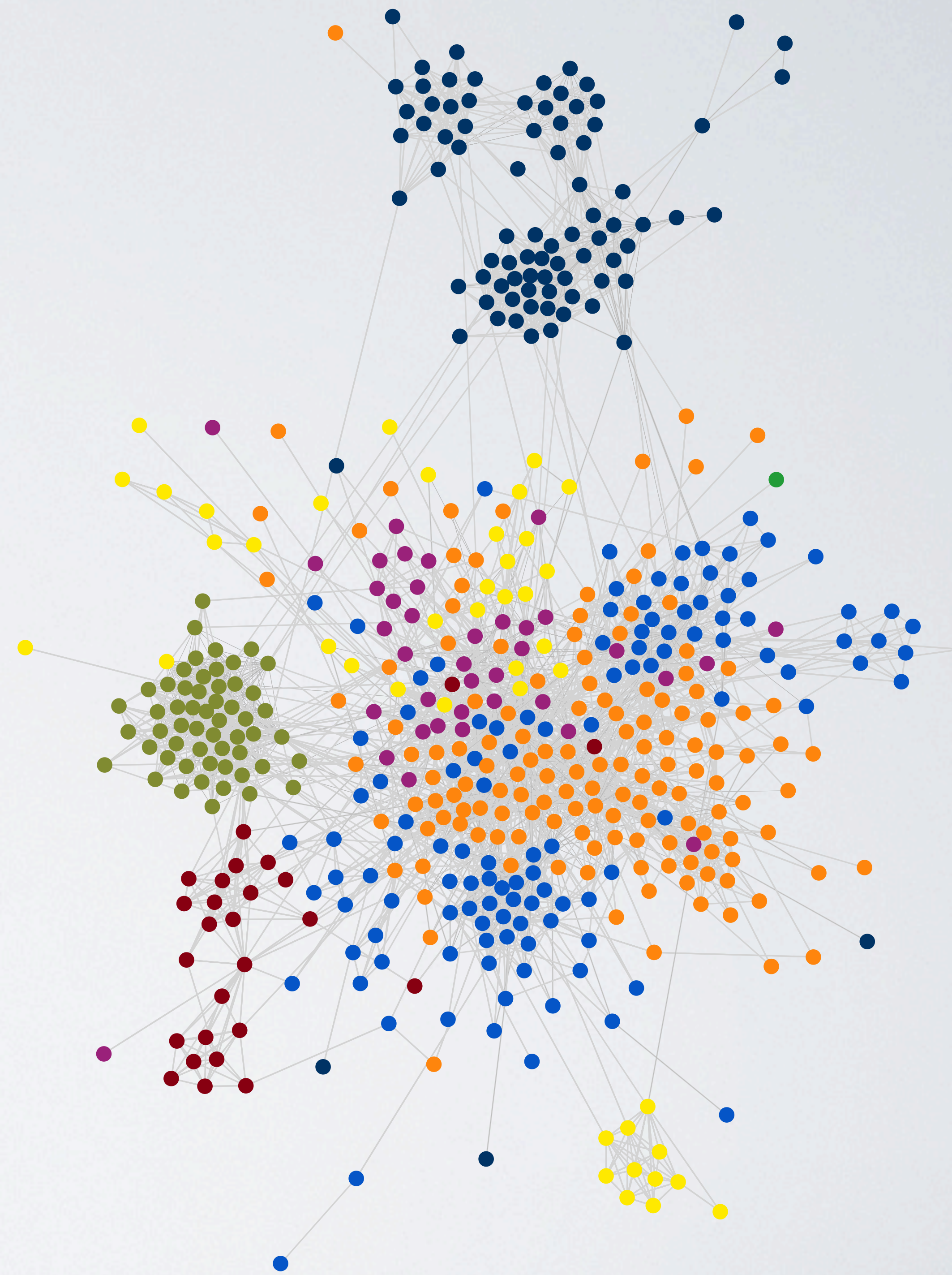
Put users into groups based on attributes

- Determine if these are communities

Need metric to rate communities

Modularity rates community strength

- Range [-1,1]
- 0 represents expected in random graph
- ≥0.25 represents community structure



Attribute communities for Rice ugrads

	Communities	Community Size			Modularity
		Min	Avg	Max	
major	65	1	23	105	0.004
matriculation year	4	95	305	398	0.259
residential college	9	130	135	142	0.385

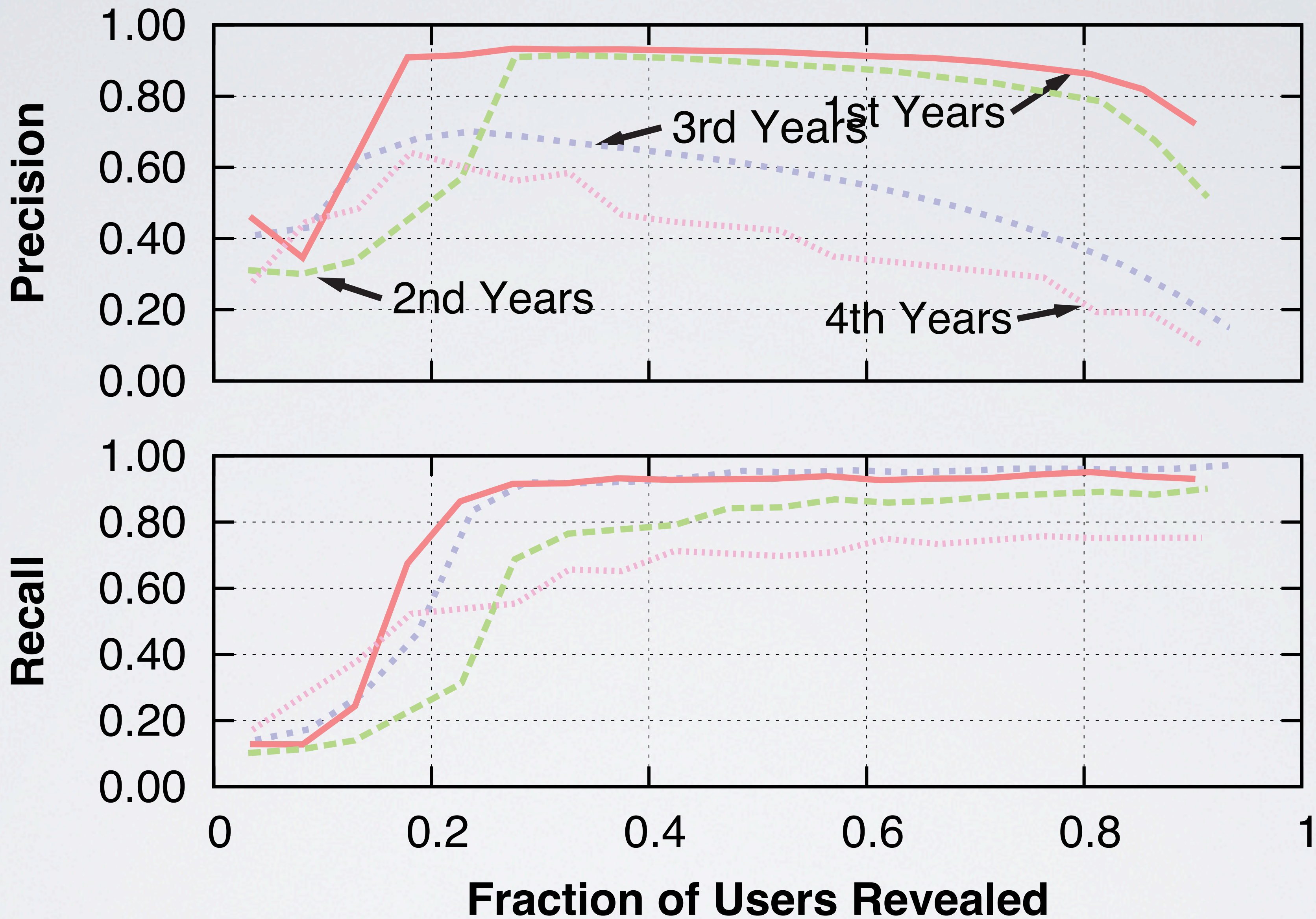
Communities based on shared college or year

- Multiple, overlapping community structures

Suggests we can build an algorithm to infer attributes

- Given a few users who share an attribute, can we guess the remaining ones?

Can we infer Rice undergrad classes?



Can infer attributes with high accuracy

- Different communities show different characteristics

Summary

Privacy an important issue in social media

- What information are users revealing without knowing it?

Demonstrated that many attributes can be inferred

- Even if user didn't provide them

Good interpretation: Can reduce burden on users

- Don't have to fill in entire profile

Bad interpretation: Can figure out attributes users don't reveal

- Privacy is a function of what friends reveal

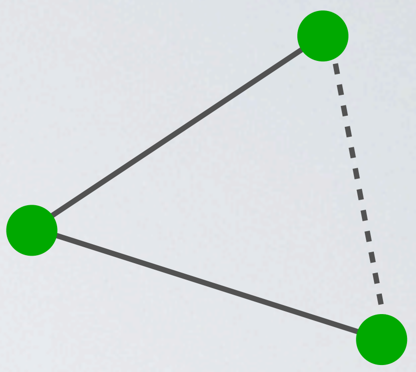
PART IV

Open questions

- or -

What should I work on?

Is Facebook changing us?



Recall strength of weak ties

- Necessary for bridging communities
- Important for information flow

Why are they weak?

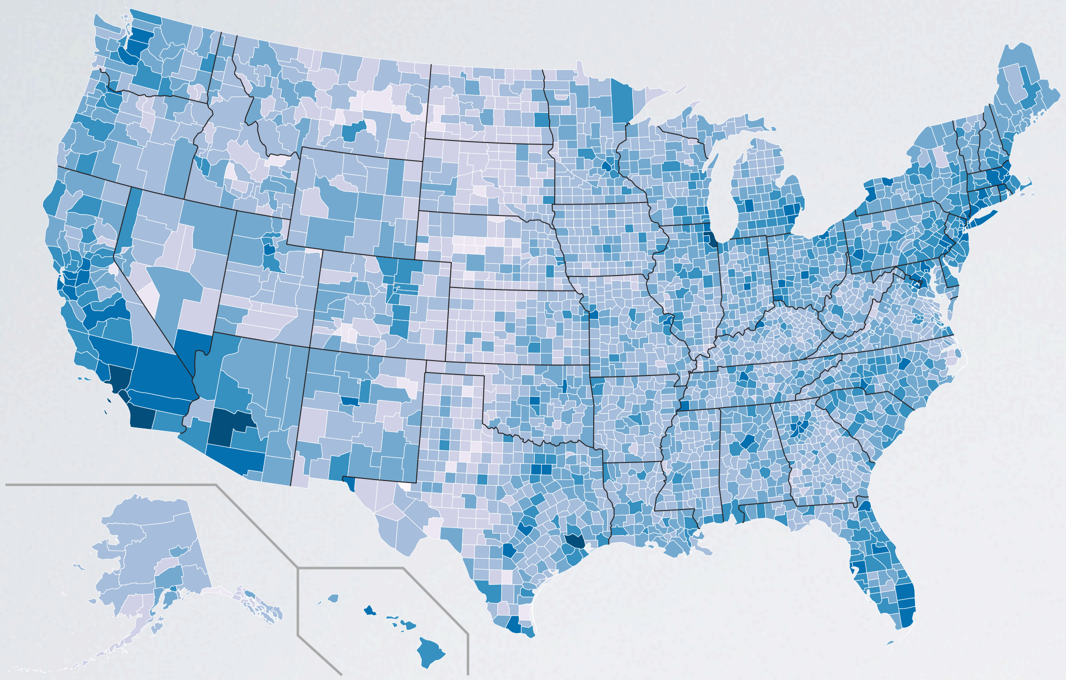
- Little interaction
- When information flows, its important

Facebook aggregates all of our weak ties

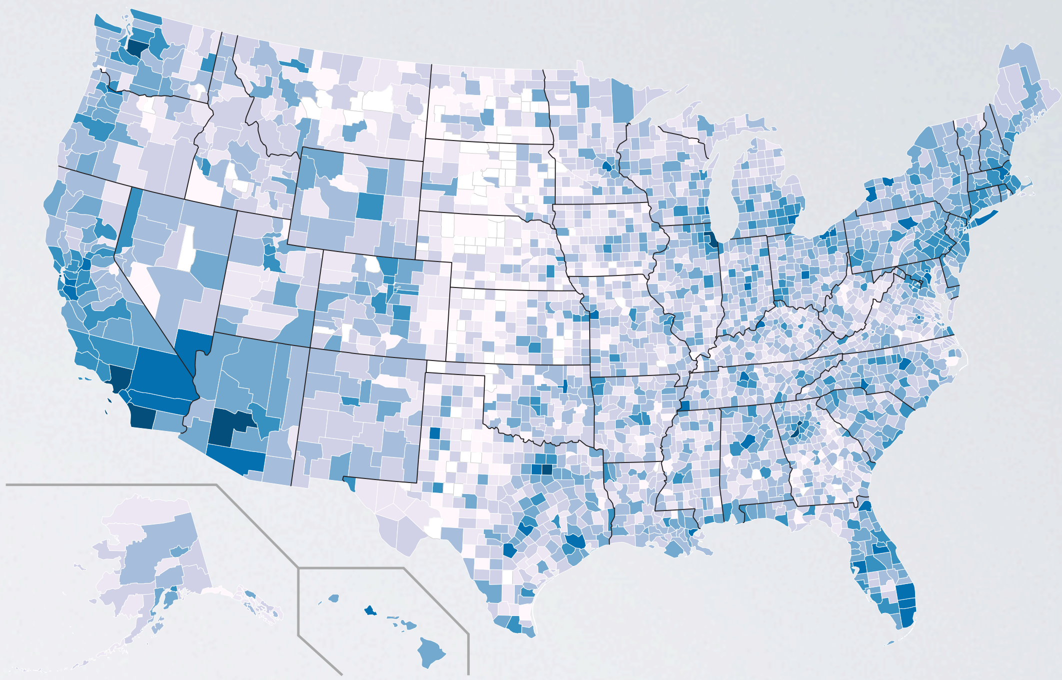
- Example: news feed often has low signal-to-noise ratio
- Mostly stories from weak ties

Is this reducing the strength of weak ties?

Is the data representative?



US Population



Twitter Users

Social media users tend to be more educated, literate, urban

- Is data obtained from social media representative?

Is there a way to correct for inherent bias?

- Can we subsample the data?

Questions?