Measurement and Analysis of Online Social Networks

Alan Mislove[†‡] Massimiliano Marcon[†] Krishna Gummadi[†]
Peter Druschel[†] Bobby Bhattacharjee[§]

[†]Max Planck Institute for Software Systems
[‡]Rice University
[§]University of Maryland

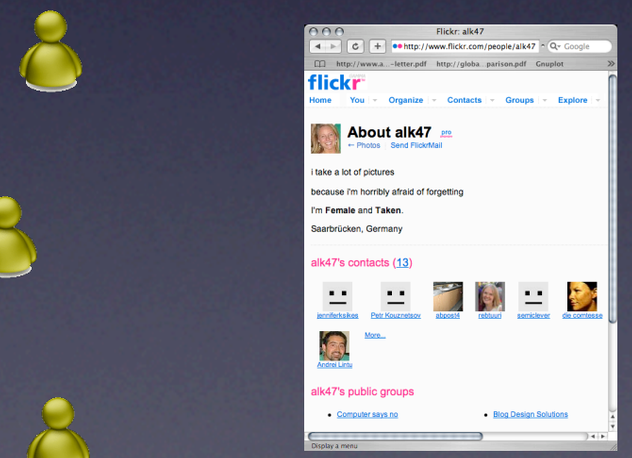
IMC 2007

What are (online) social networks?

- Social networks are graphs of people
  - Graph edges connect friends

Social Network

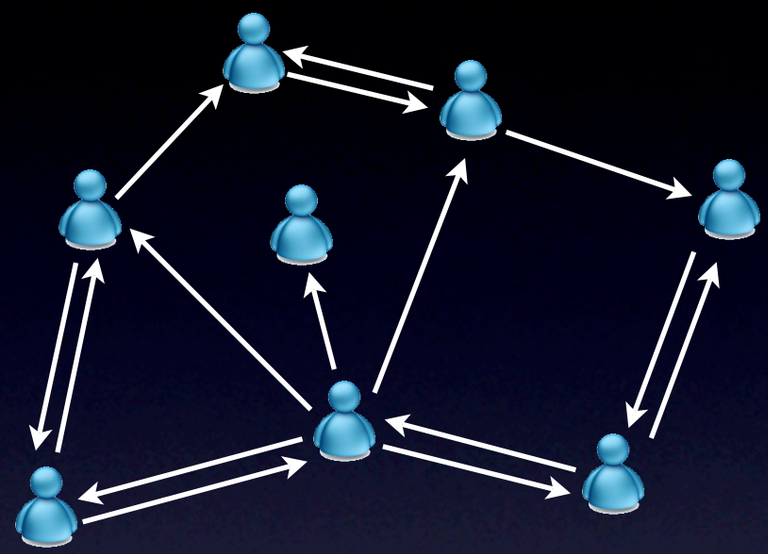
- Online social networking
  - Social network hosted by a Web site
  - Friendship represents shared interest or trust
  - Online friends may have never met



Online Social Network

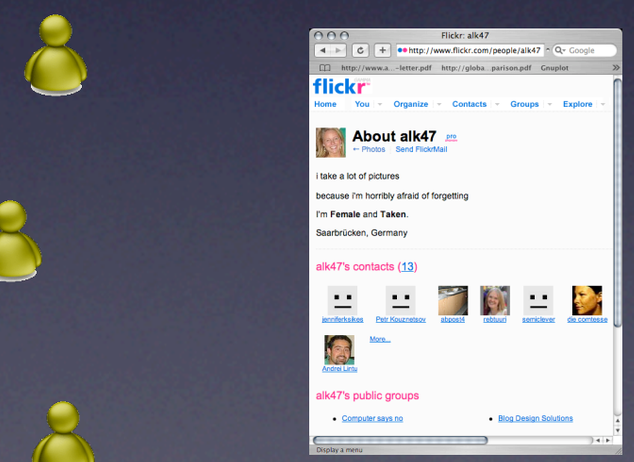
What are (online) social networks?

- Social networks are graphs of people
  - Graph edges connect friends



Social Network

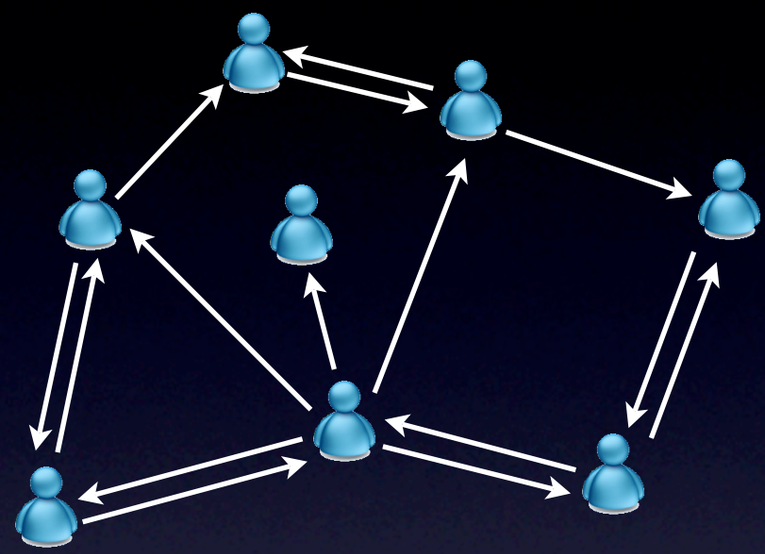
- Online social networking
  - Social network hosted by a Web site
  - Friendship represents shared interest or trust
  - Online friends may have never met



Online Social Network

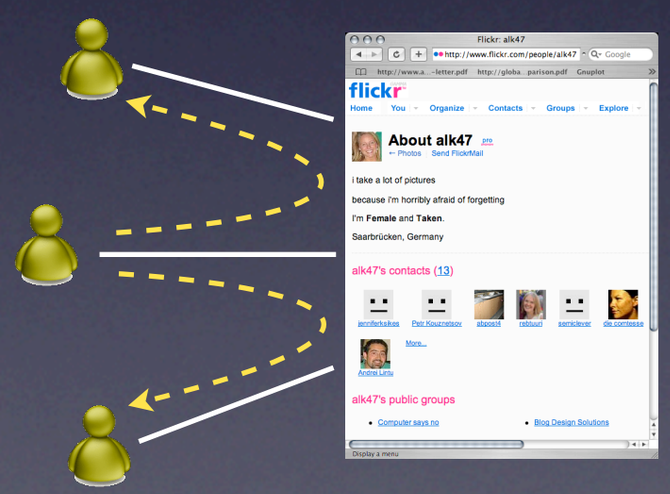
What are (online) social networks?

- Social networks are graphs of people
  - Graph edges connect friends



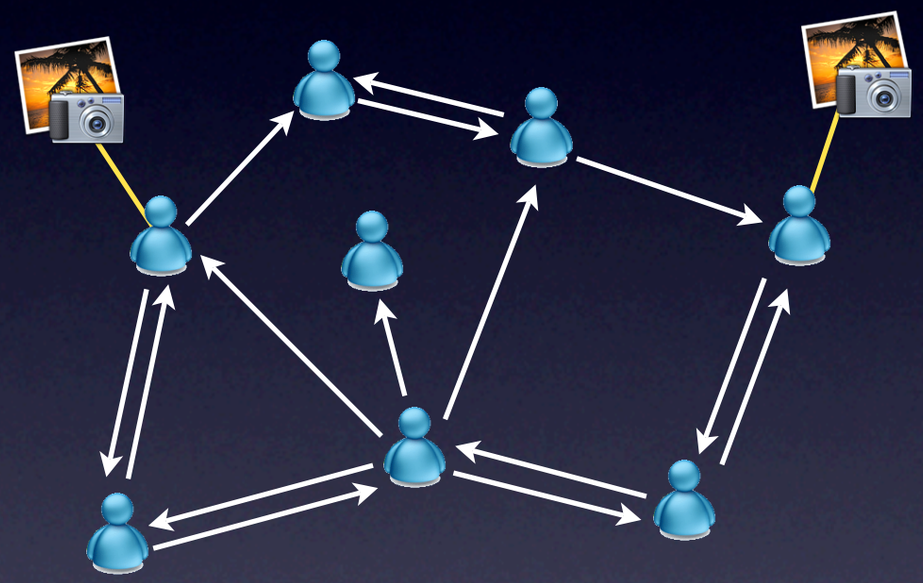
Social Network

- Online social networking
  - Social network hosted by a Web site
  - Friendship represents shared interest or trust
  - Online friends may have never met



Online Social Network

What are online social networks used for?



- Popular way to connect, share content
  - Photos (Flickr), videos (YouTube), blogs (LiveJournal), profiles (Orkut)
  - Orkut (60 M), LiveJournal (5 M)
- Content organized with user-user links
  - Akin to Web's page-page links
  - Social network *structure* influences how content is shared

This work

- Presents large-scale measurement study and analysis 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

- Our goals are two-fold:
  - Measure online social networks at scale
  - Understand static structural properties

Why study social network structure?

- Guide designers of future systems
  - Trust relationships suggest new reasoning about trust
  - Shared interest suggests new ways of structuring information

- Trust can be used to solve security problems
  - Multiple identity attacks: SybilGuard [SIGCOMM'06]
  - Spam: RE [NSDI'06]

- Shared interest can improve content location
  - Web search: PeerSpective [HotNets'06]

- Understanding network structure is necessary first step

Rest of the talk

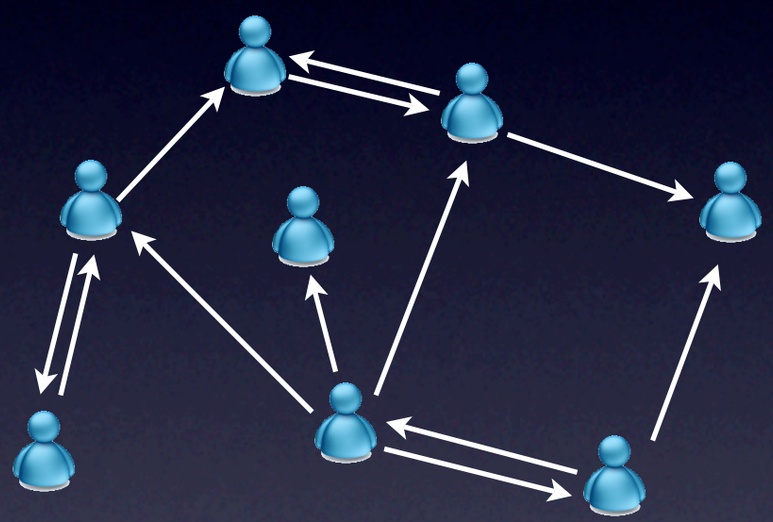
- Measuring social networks at scale
- Analyzing structural properties

Overview: Measuring online 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

Overview: Measuring online 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

Challenges faced

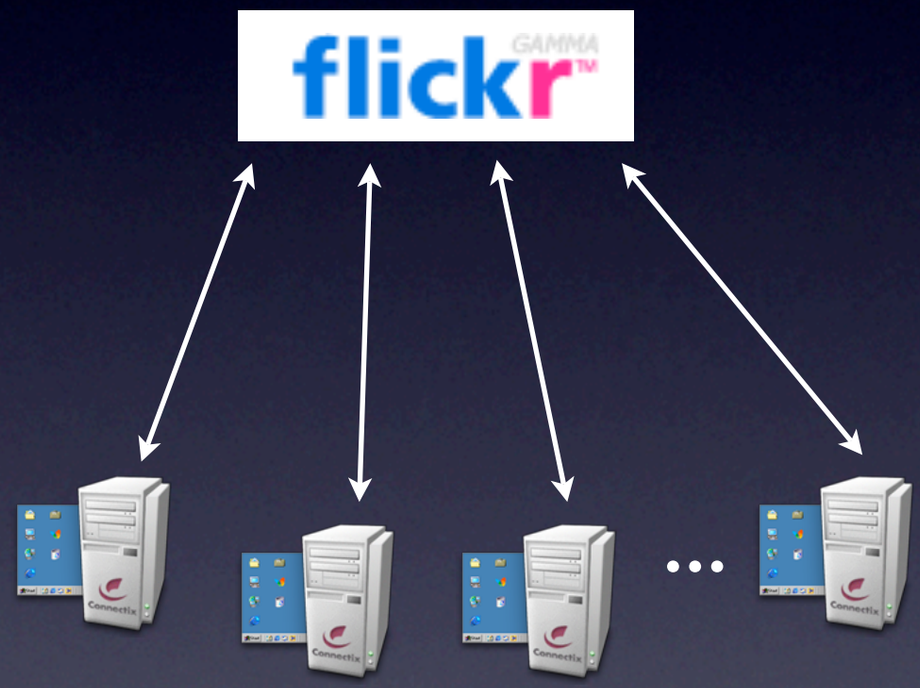
- Obtaining data using crawling presents unique challenges
- Crawling quickly
  - Underlying social networks changing rapidly
    - Consistent snapshot hard to get
    - Need to complete the crawl quickly
- Crawling completely
  - Social networks aren't necessarily connected
    - Some users have no links, or small clusters
    - Need to estimate the crawl coverage

How fast could we crawl?

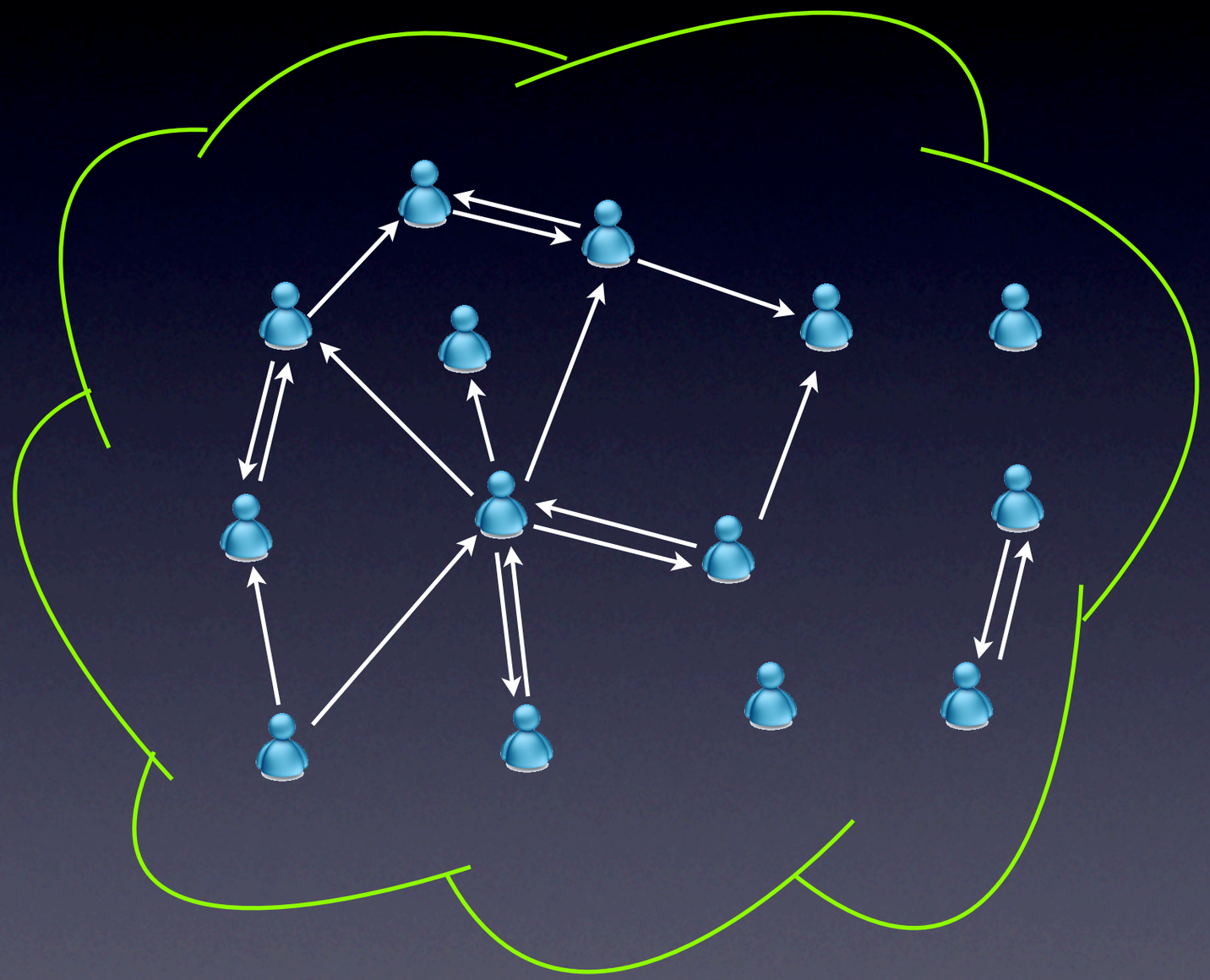
- Crawled using cluster of 58 machines
  - Used APIs where available
  - Otherwise, used screen scraping

- Crawls took varying times
  - Flickr, YouTube: 1 day
  - LiveJournal: 3 days
  - Orkut (partial): 39 days

- Crawls subject to rate-limiting
  - Discovered appropriate rates

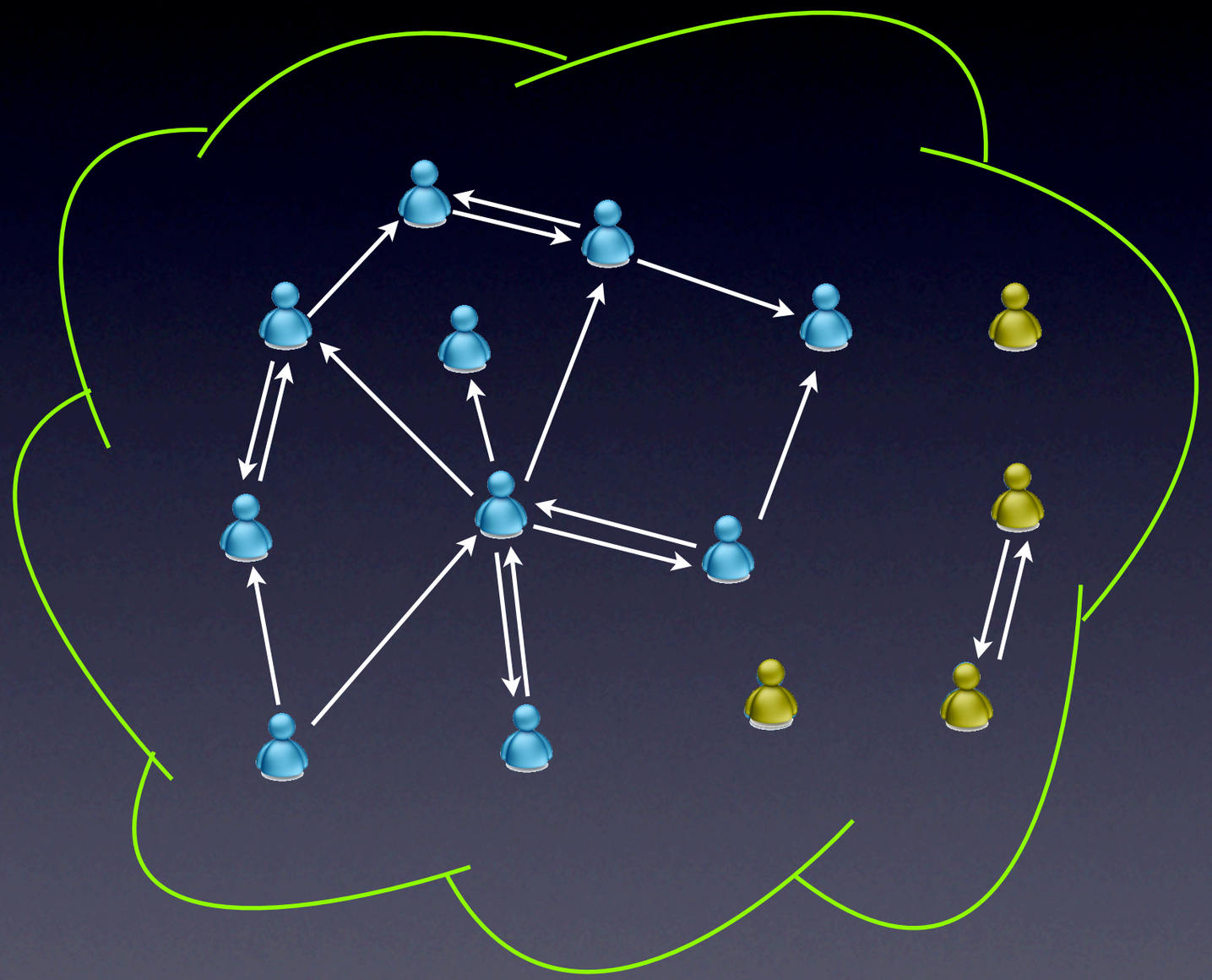


How much could we crawl?



- Users don't necessarily form single WCC
  - Disconnected users
- Estimate coverage by selecting random users
  - After crawl, determine fraction of users covered
- Networks tend to have one giant WCC

How much could we crawl?



- Users don't necessarily form single WCC
  - Disconnected users
- Estimate coverage by selecting random users
  - After crawl, determine fraction of users covered
- Networks tend to have one giant WCC

Evaluating coverage: Flickr

- Obtained random users by guessing usernames (##########@N00)
- Fraction of disconnected users is 73%
- But, disconnected users have very low degree
  - 90% have no outgoing links, remaining 10% have few links
- Summary:
  - Covered 27% of user population, but remaining users have very few links

Evaluating coverage: LiveJournal

- Obtained random users using special URL
  - `http://www.livejournal.com/random.bml`
- Fraction of disconnected users is only 5%

- Summary:
  - Crawl covered 95% of user population

Evaluating coverage: Orkut

- At time of crawl, Orkut was fully connected
  - But, we ended crawl early
- How representative is our sub-crawl?
  - Performed multiple crawls from different seeds
  - Obtained random seed users using maximum-degree sampling
- Properties consistent across smaller crawls
- Summary:
  - Sub-crawl of user population, but likely representative of similarly sized subcrawls

Evaluating coverage: YouTube

- Could not obtain random users
  - Usernames user-specified strings
  - Not fully connected (could not use maximum-degree sampling)

- Unable to find estimate of user population

- Summary:
  - Unable to estimate fraction of users covered

Outline

- ~~Measuring social networks at scale~~
- Analyzing structural properties

Network structure questions

- Want to examine structural properties

- Which users have the links?
  - Even distribution of links, or is it skewed?

- Are there a few nodes holding the network together?
  - Or, is the network robust?

- How do social networks differ from known networks?
  - Such as the Web

High-level data characteristics

	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

High-level data characteristics

	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

High-level data characteristics

	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

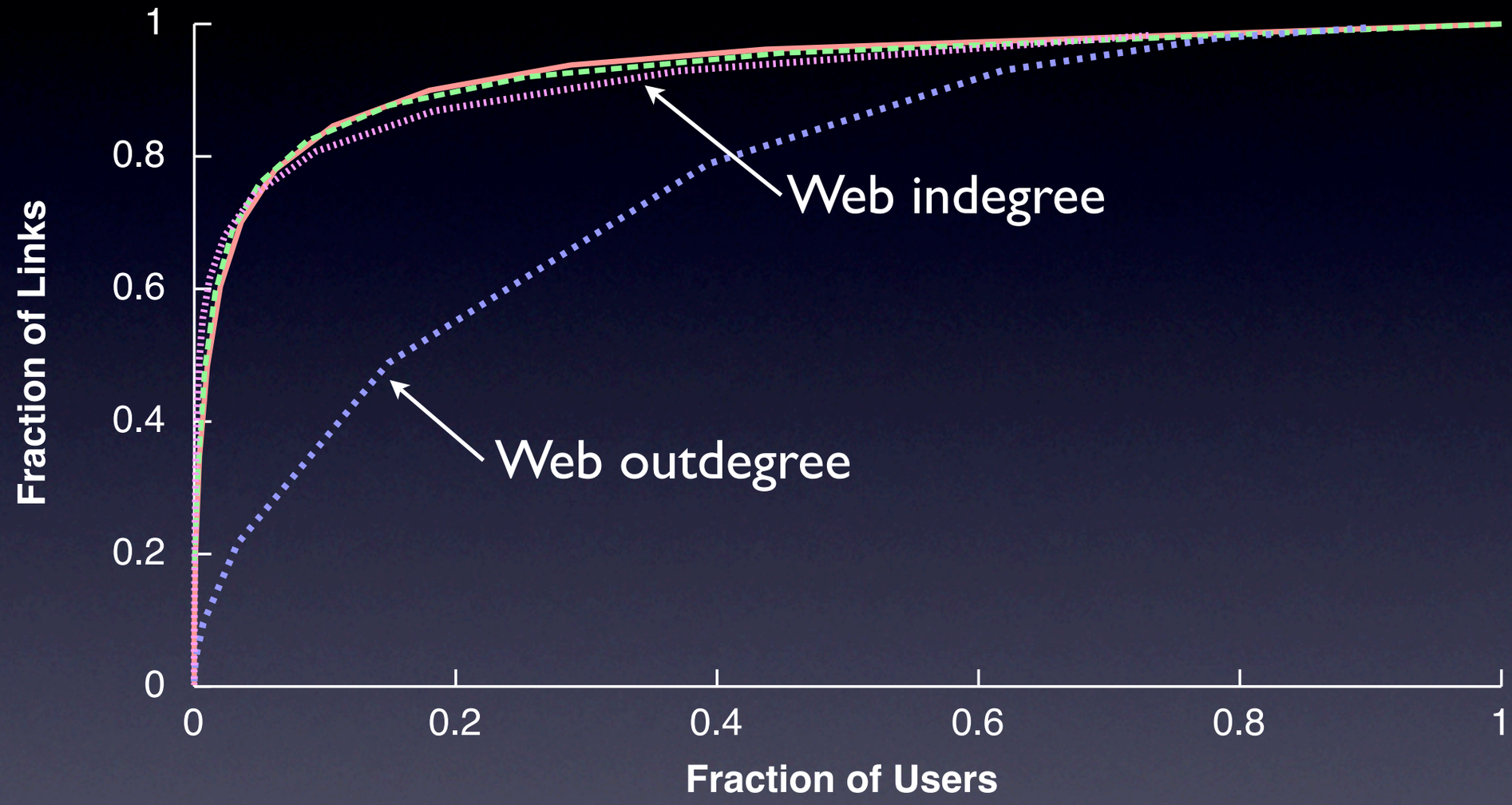
Are online social networks power-law?

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

- Estimated coefficients with maximum likelihood testing
  - Flickr, LiveJournal, YouTube have good K-S goodness-of-fit
  - Orkut deviates due to partial crawl

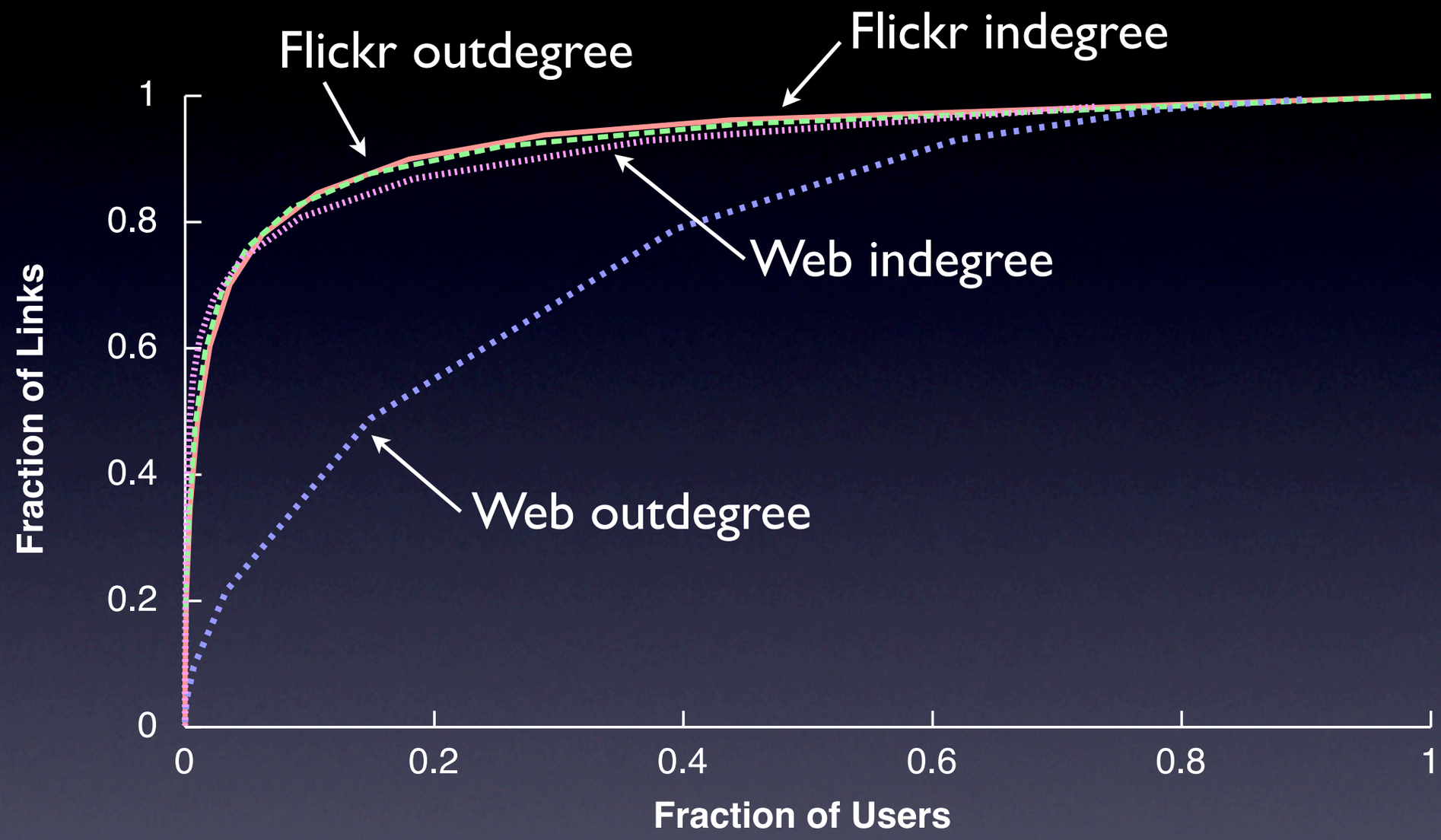
- Similar coefficients imply a similar distribution of in/outdegree
  - Unlike Web [INFOCOMM'99]

How are the links distributed?



- Distribution of indegree and outdegree is similar
  - Underlying cause is *link symmetry*

How are the links distributed?



- Distribution of indegree and outdegree is similar
  - Underlying cause is *link symmetry*

Link symmetry

- Social networks show high level of link symmetry
  - Links in most networks are directed

	Flickr	LiveJournal	Orkut	YouTube
Symmetric Links				

- High symmetry increases network connectivity
  - Reduces network diameter

Link symmetry

- Social networks show high level of link symmetry
  - 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

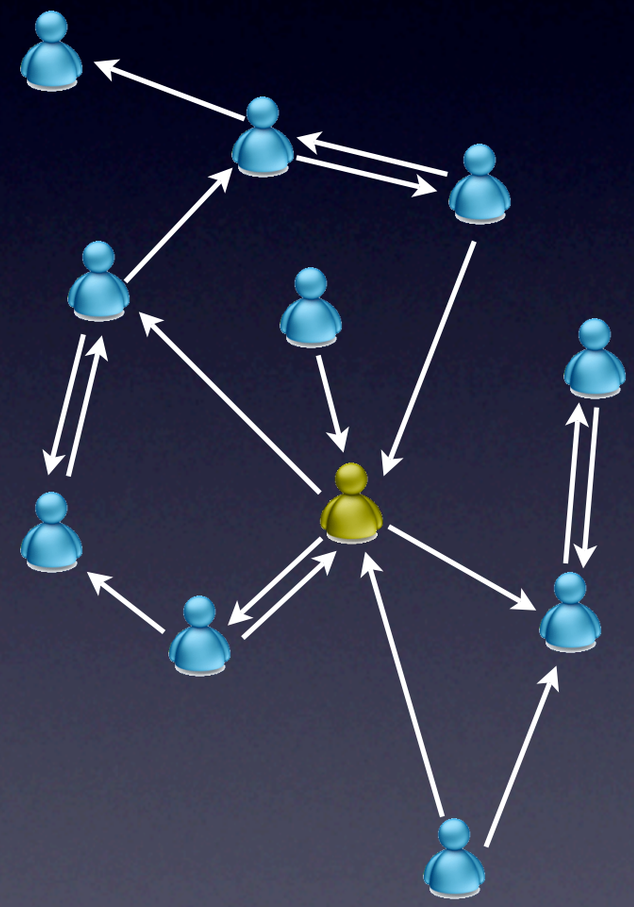
- Unlike other complex networks, such as the Web
  - Sites like **`cnn.com`** receive much links more than they give

- 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? [JCN'06]
  - A bowtie, like the Web? [WWW'00]

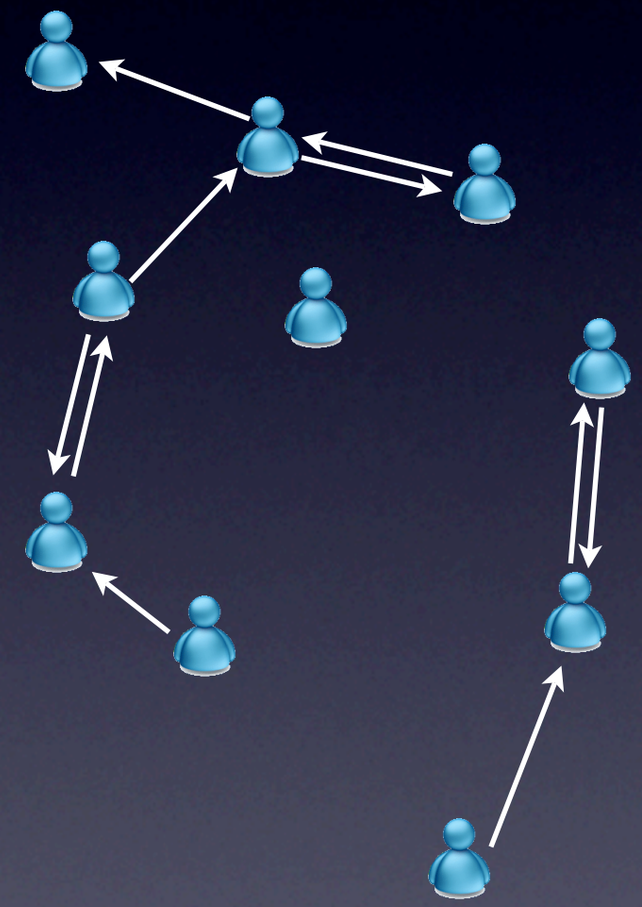


- In particular, is there a *core* of the network?
  - Core is a (minimal) connected component
  - Removing core disconnects remaining nodes

- Approximate core detection by removing high-degree nodes

Complex network structure

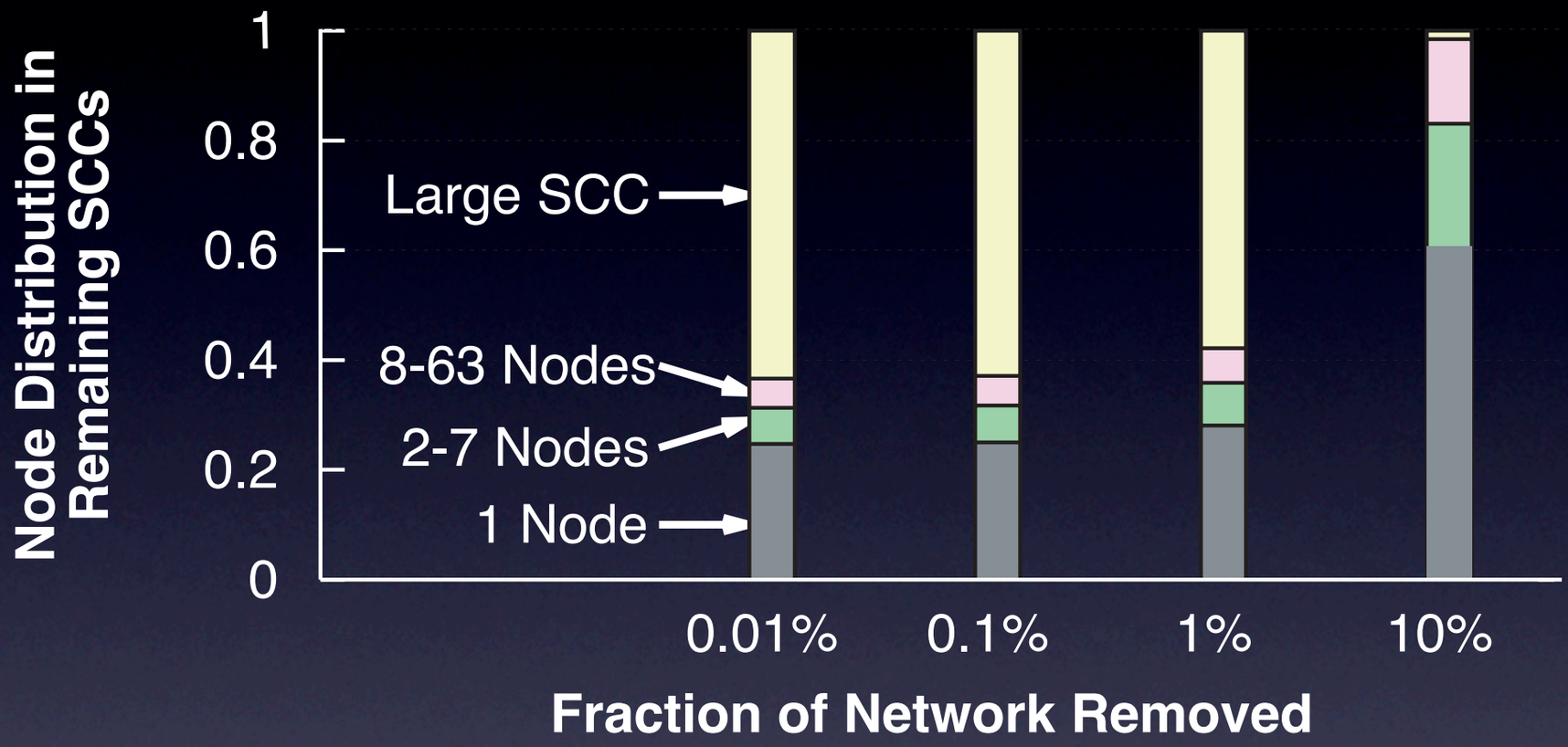
- What is the high-level structure of online social networks?
  - A jellyfish, like the Internet? [JCN'06]
  - A bowtie, like the Web? [WWW'00]



- 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

Does a core exist?

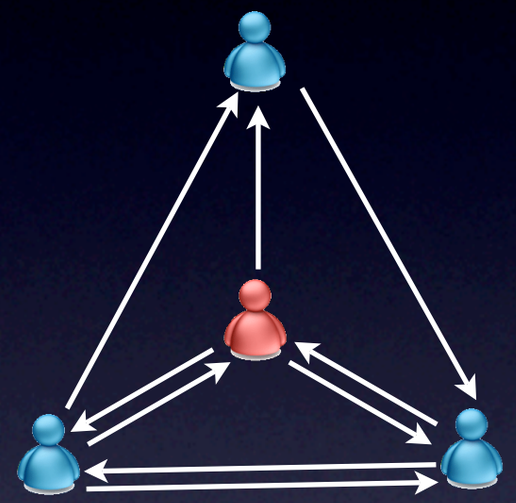


- Yes, networks contain core consisting of 1-10% of nodes
  - Removing core disconnects other nodes
- What about remaining nodes (the fringe)?

Clustering

- Clustering coefficient *C* is a metric of cliquishness

$$C = \frac{\text{Number of links between friends}}{\text{Number of links that could exist}}$$



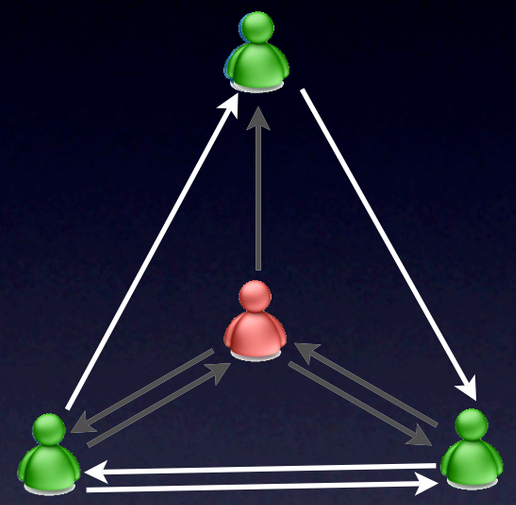
- Online social networks are tightly clustered
  - 10,000 times more clustered than random graphs
  - 5-50 times more clustered than random power-law graphs

- How is the network clustered?

Clustering

- Clustering coefficient *C* is a metric of cliquishness

$$C = \frac{\text{Number of links between friends}}{\text{Number of links that could exist}}$$



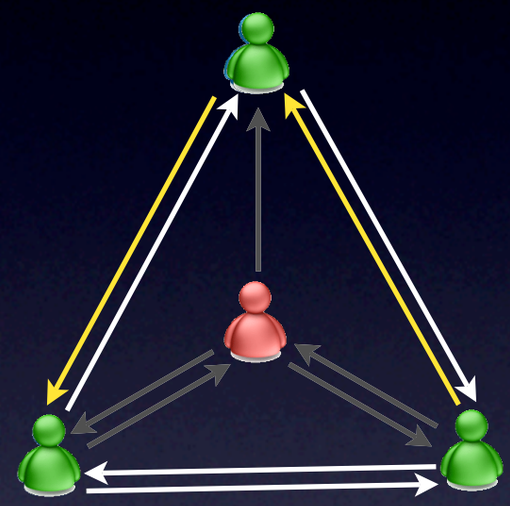
- Online social networks are tightly clustered
  - 10,000 times more clustered than random graphs
  - 5-50 times more clustered than random power-law graphs

- How is the network clustered?

Clustering

- Clustering coefficient *C* is a metric of cliquishness

$$C = \frac{\text{Number of links between friends}}{\text{Number of links that could exist}}$$

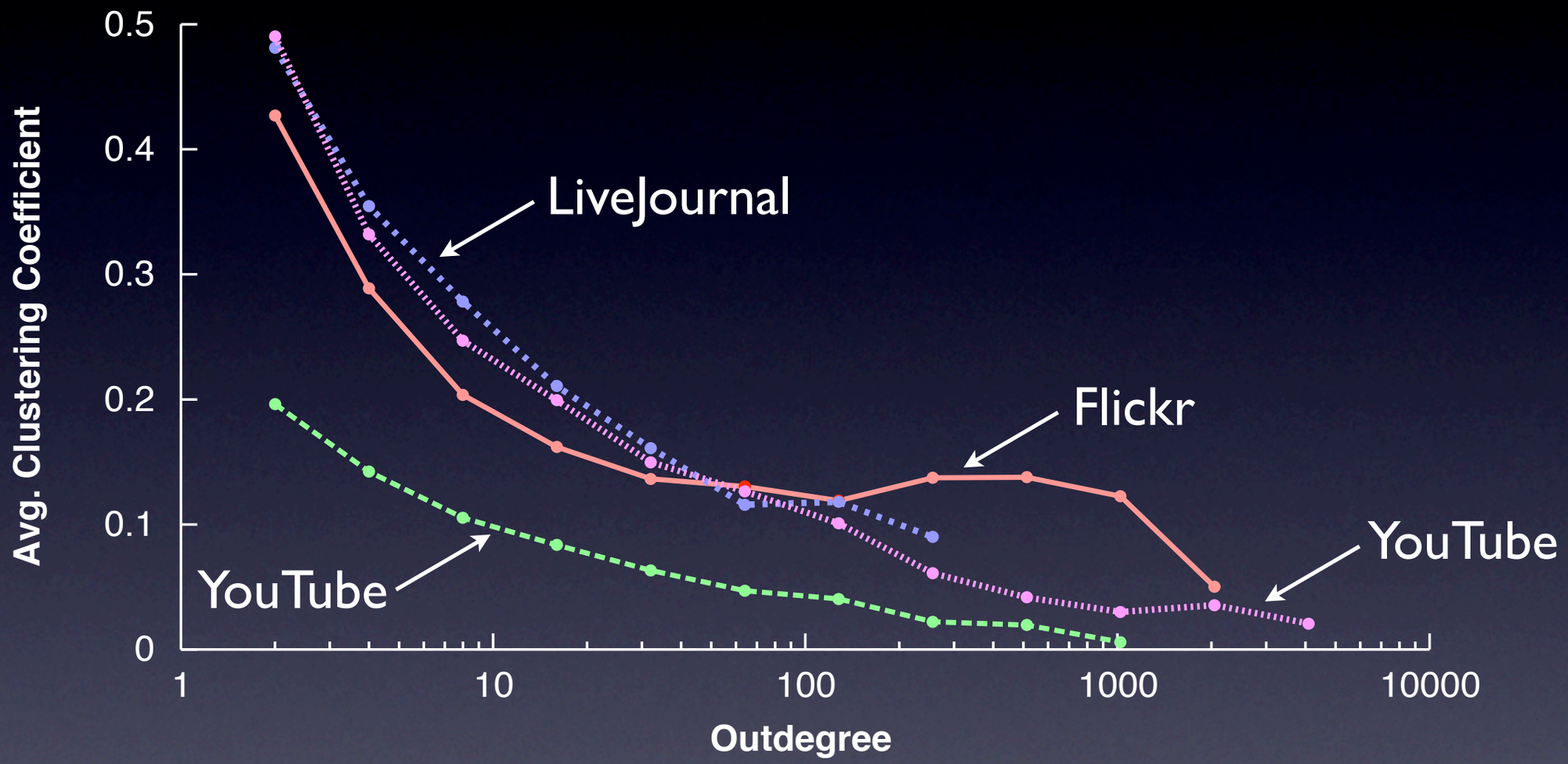


C = 0.66

- Online social networks are tightly clustered
  - 10,000 times more clustered than random graphs
  - 5-50 times more clustered than random power-law graphs

- How is the network clustered?

Are the fringes more clustered?



- Low-degree users show high degree of clustering
  - Networks are small-world, may be scale-free

Implications of network structure

- 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

Summary

- Presented first large-scale study of multiple online social networks
- Outlined challenges with crawling large networks
  - Able to overcome challenges with multiple sites
- Analyzed and compared network structure
  - Multiple networks have similar, unique characteristics
- Data sets are available to researchers
  - Many already using data (12 research groups, including sociologists!)

http://socialnetworks.mpi-sws.org