UNIVERSITY OF CAMBRIDGE
800 YEARS
1209 ~ 2009

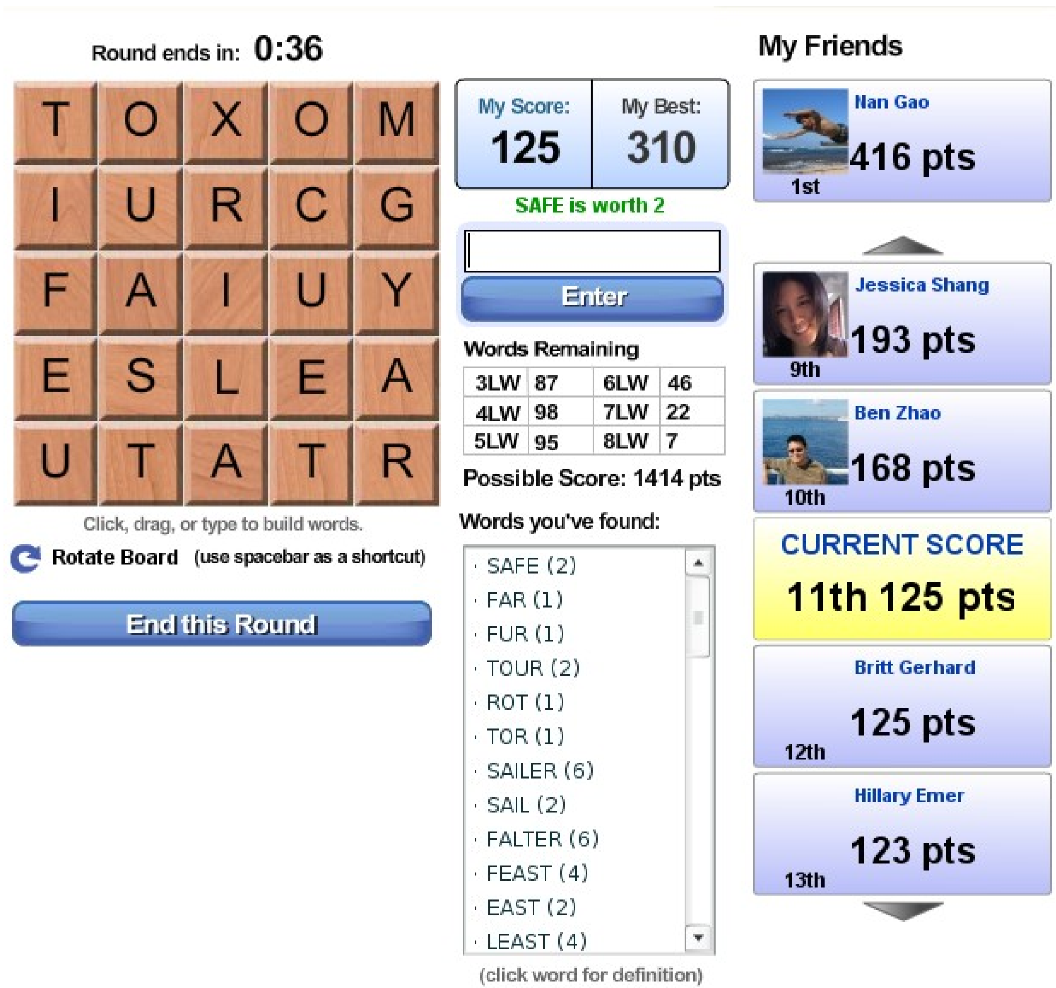
Privacy Aspects of Social Graphs

Joseph Bonneau

Stanford Security Seminar, July 14 2009

University of Cambridge Computer Laboratory

Social Context And The Web



Everything's Better With Friends...

- "Hyper-presence" of friends
- "networked public spaces"
- All web activity will have social context

Facebook Is Becoming A Second Internet...

Function	Internet version	Facebook version
Page Markup	HTML, JavaScript	FBML
DB Queries	SQL	FBQL
Email	SMTP	FB Mail
Forums	Usenet, etc.	FB Groups
Instant Messages	XMPP	FB Chat
News Streams	RSS	FB Stream
Authentication	OpenID	FB Connect
Photo Sharing	Flickr, etc.	FB Photos
Video Sharing	YouTube, etc.	FB Video
Blogging	Blogger, etc.	FB Notes
Microblogging	Twitter, etc.	FB Status Updates
Micropayment	Peppercoin, etc.	FB Points
Event Planning	E-Vite	FB Events
Classified Ads	craigslist	FB Marketplace

Parallel Trend: The Internet is Becoming Social

"Given sufficient funding, all web sites expand in functionality until users can add each other as friends"

“Traditional” Social Network Analysis

- Performed by sociologists, anthropologists, etc. since the 70's
- Use data carefully collected through interviews & observation
  - Typically < 100 nodes
  - Complete knowledge
  - Links have consistent meaning
- All of these assumptions fail badly for online social network data

Traditional Graph Theory

- Nice Proofs
- Tons of definitions
- Ignored topics:
  - Large graphs
  - Sampling
  - Uncertainty

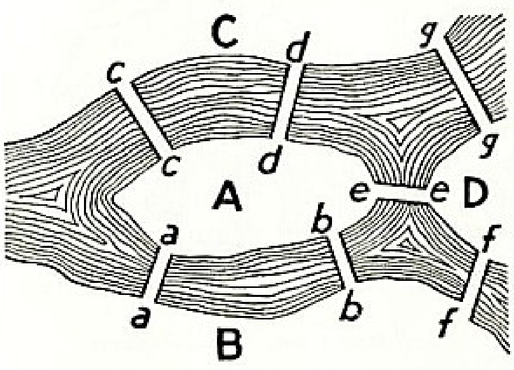
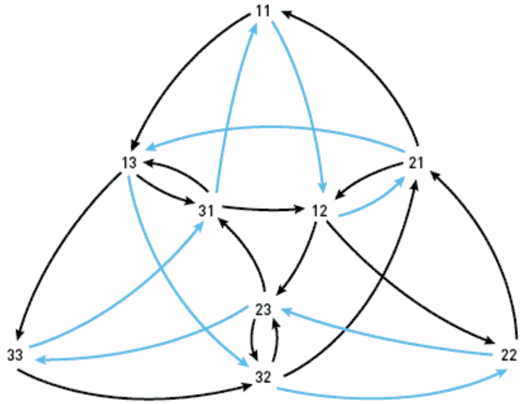


FIGURE 98. *Geographic Map: The Königsberg Bridges.*

HAMILTON CYCLE ON DE BRUIJN GRAPH



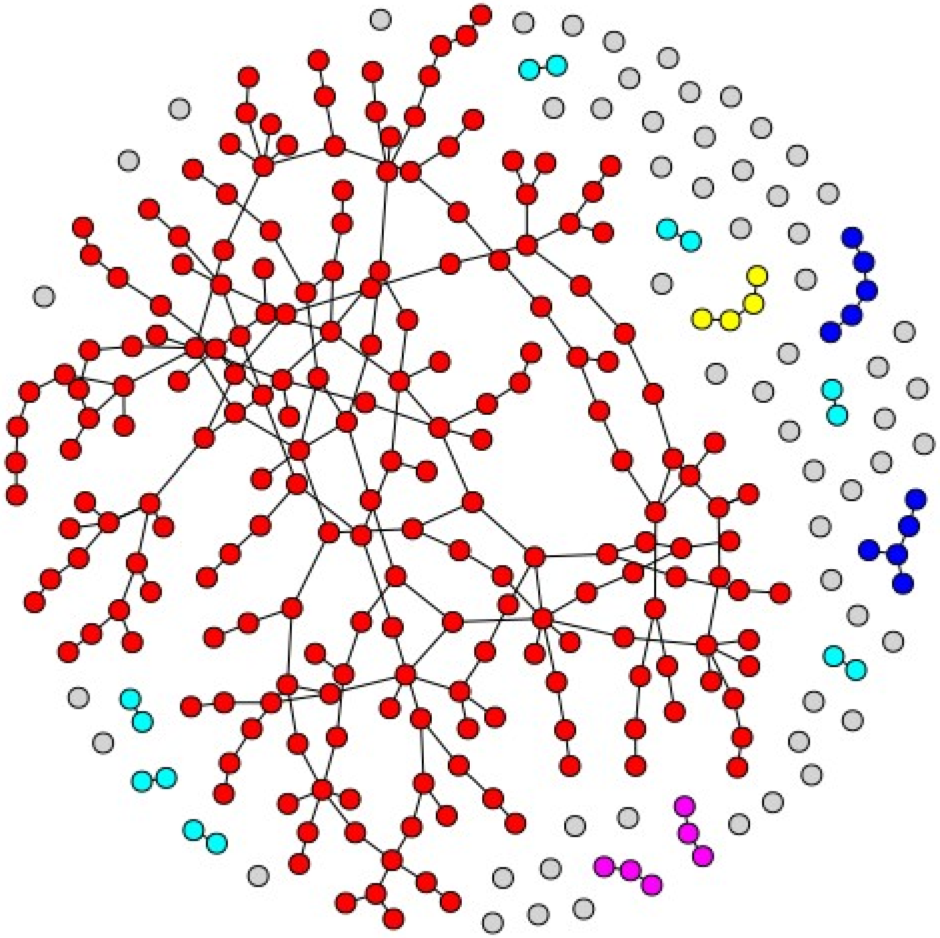
Models Of Complex Networks From Math & Physics

Many nice models

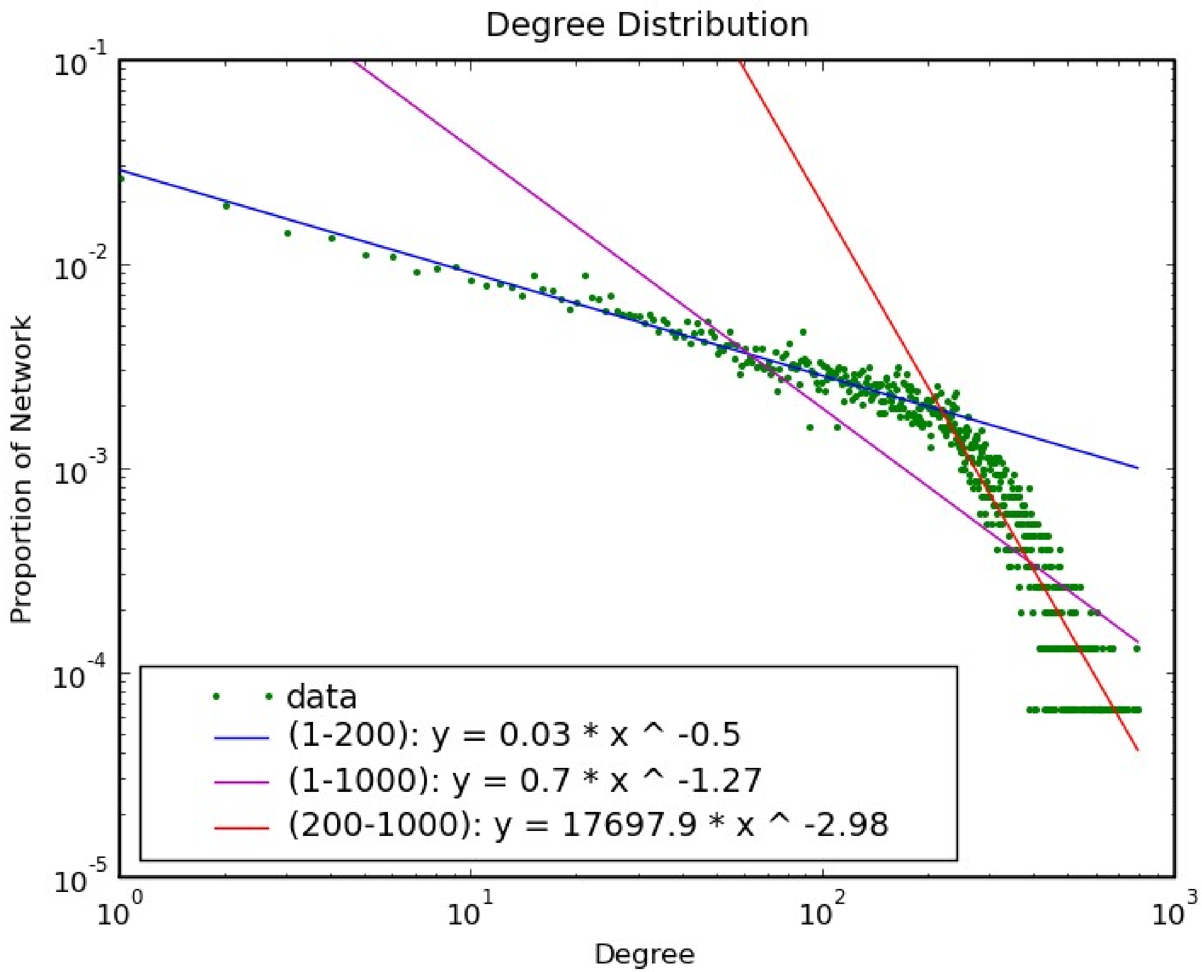
- Erdos-Renyi
- Watts-Strogatz
- Barabasi-Albert

Social Networks properties:

- Power-law
- Small-world
- High clustering coefficient



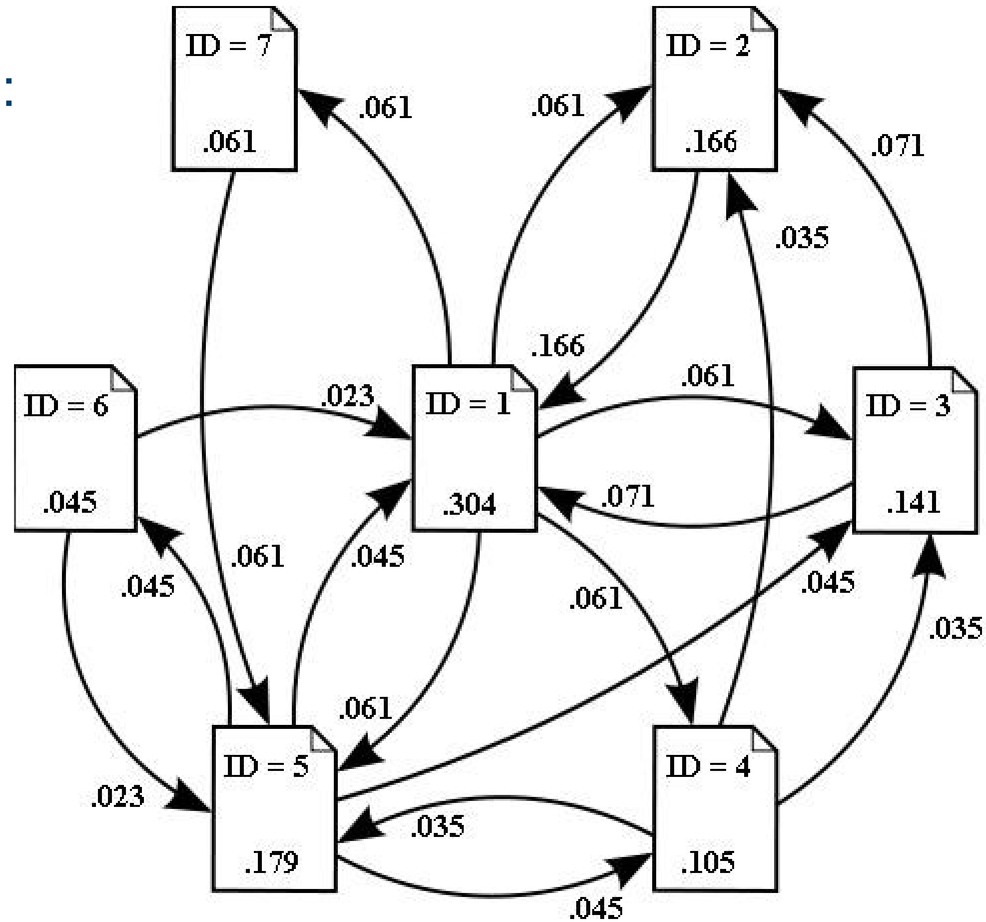
Real social graphs are complicated!



When In Doubt, Compute!

We do know many graph algorithms:

- Find important nodes
- Identify communities
- Train classifiers
- Identify anomalous connections



Major Privacy Implications!

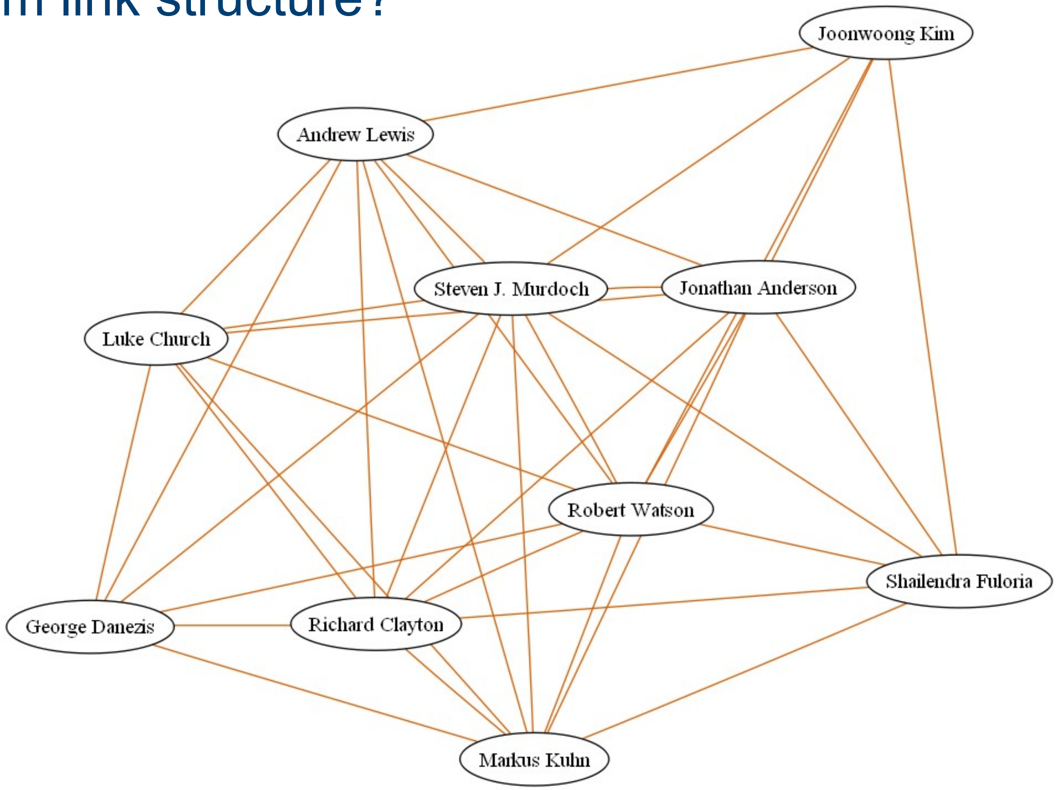
Privacy Questions

- What can we infer purely from link structure?

Privacy Questions

- What can we infer purely from link structure?

  A surprising amount!
- Popularity
- Centrality
- Introvert vs. Extrovert
- Leadership potential



Privacy Questions

- If we know nothing about a node but it's neighbours, what can we infer?

Privacy Questions

- If we know nothing about a node but it's neighbours, what can we infer?

  A lot!

- Gender
- Political Beliefs
- Location
- Breed?

Privacy Questions

- Can we anonymise graphs?

Privacy Questions

- Can we anonymise graphs?

  Not easily...

- Seminal result by Backstrom et al.: Attack of attack needs just 7 nodes
- Can do even better given user's complete neighborhood
- Also results for correlating users across networks
- Developing line of research...

Privacy Questions

- What can we infer if we “compromise” a fraction of nodes?

Privacy Questions

- What can we infer if we “compromise” a fraction of nodes?

  A lot...

- Common theme: small groups of nodes can see the rest
  - Danezis et al.
  - Nagaraja
  - Korolova et al.
  - Bonneau et al.

Privacy Questions

- Can we defend against crawling in a sound way?

  Work in progress!

Privacy Questions

- What if we get a subset of neighbours for all nodes?

Privacy Questions

- What if we get a subset of *k* neighbours for all nodes?

  Emerging question for many social graphs

  - Facebook and online SNS
  - Mobile SNS

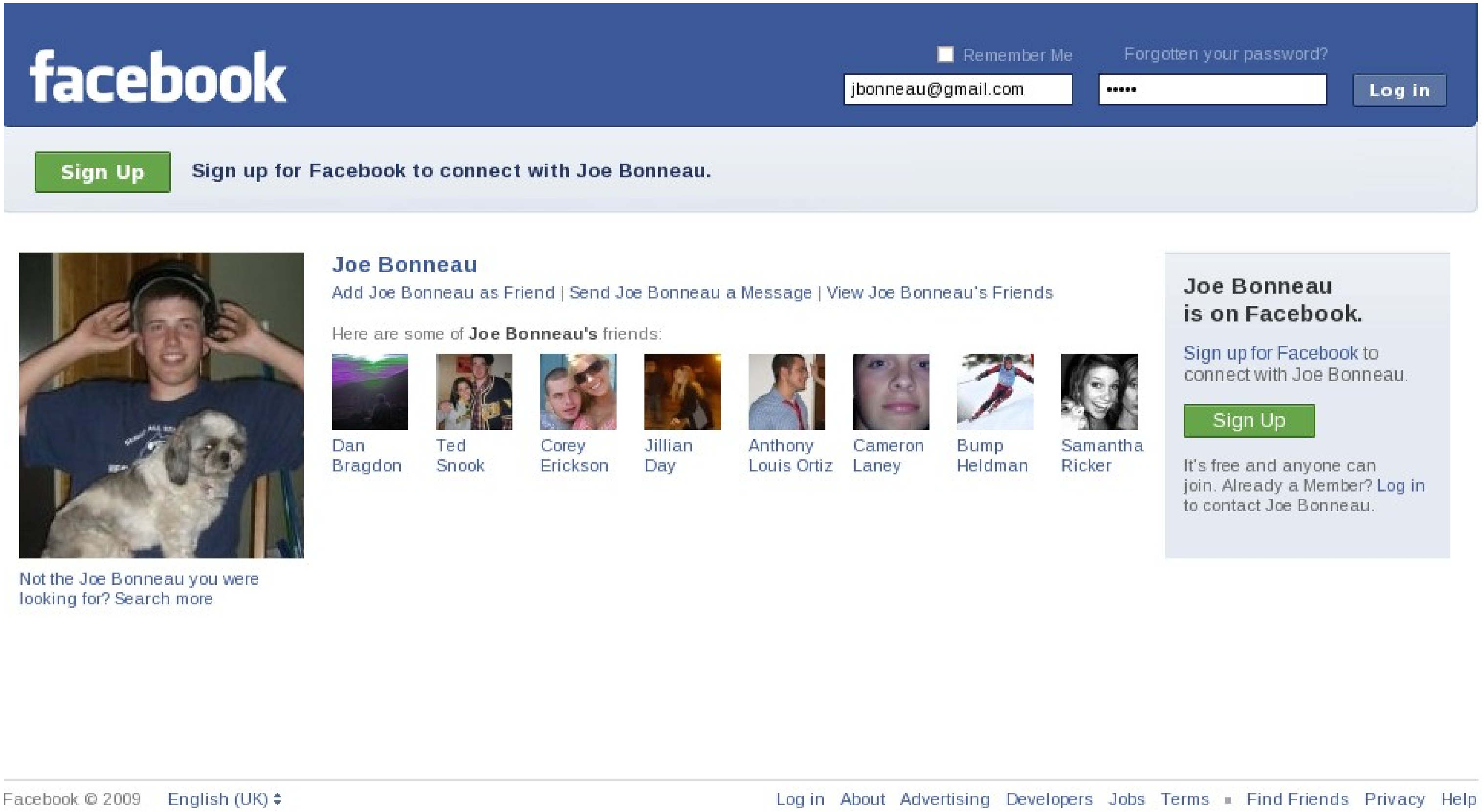
A Quietly Introduced Feature...

Public Search Listings, Sep 2007

Public Search Listings

- Unprotected against crawling
- Indexed by search engines
- Opt out—but most users don't know it exists!

Utility



Entity Resolution

Utility

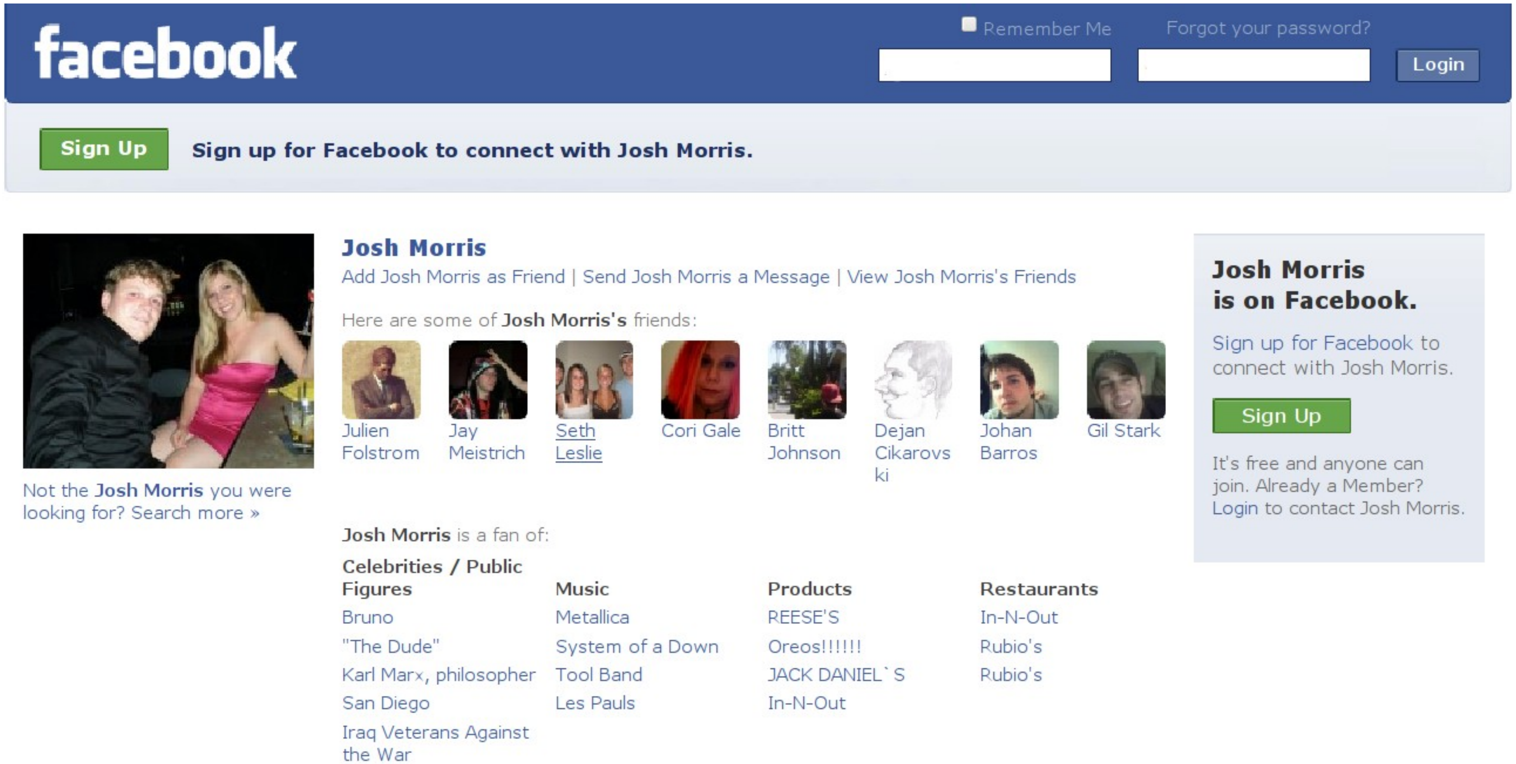
Promotion via Network Effects

Legal Status

"Your name, network names, and profile picture thumbnail will be available in search results across the Facebook network and those limited pieces of information may be made available to third party search engines. This is primarily so your friends can find you and send a friend request."

-Facebook Privacy Policy

Legal Status



Much More Info Now Included...

Legal Status

Advocates of Communism
Global

Basic Info

Type: Common Interest - Politics

Description: The working class has nothing to lose but their chains. They have the world to win.

We have seen above that the first step in the revolution by the working class is to raise the proletariat to the position of ruling class to win the battle of democracy.

The proletariat will use its political supremacy to wrest, by degree, all capital from the bourgeoisie, to centralize all instruments of production in the hands of the state, i.e., of the proletariat organized as the ruling class; and to increase the total productive forces as rapidly as possible.

Of course, in the beginning, this cannot be effected except by means of despotic inroads on the rights of property, and on the conditions of bourgeois production; by means of measures, therefore, which appear economically insufficient and untenable, but which, in the course of the movement, outstrip themselves, necessitate further inroads upon the old social order, and are unavoidable as a means of entirely revolutionizing the mode of production.

These measures will, of course, be different in different countries.

Nevertheless, in most advanced countries, the following will be pretty generally applicable.

1. Abolition of property in land and application of all rents of land to public purposes.

2. A heavy progressive or graduated income tax.

Group Type

This is an open group. Anyone can join and invite others to join.

Officers

Sim
Party Philosopher

Nils
Questioner of Party Authority

Aaron
Official Representative of Anti-Revisionist Socialism

Aziz
Official Representative of U.C.Y

William
Official Representative of Moderate Trotskyist Party

Pmk
Official Representative of Communist Revolution Party

Will
Official Representative of Utopia Party

Adem
Vice Representative of the Downtrodden

Members

Displaying 8 of 3,513 members

Logan Akosua James Thomas Raph Austin Irvin Lahiru

Public Group Pages Recently Added

Obvious Attack

- Initially returned new friend set on refresh
- Can find all *n* friends in O(*n*·log *n*) queries
  - The Coupon Collector's Problem
  - For 100 Friends, need 65 page refreshes
- As of Jan 2009, friends fixed per IP address

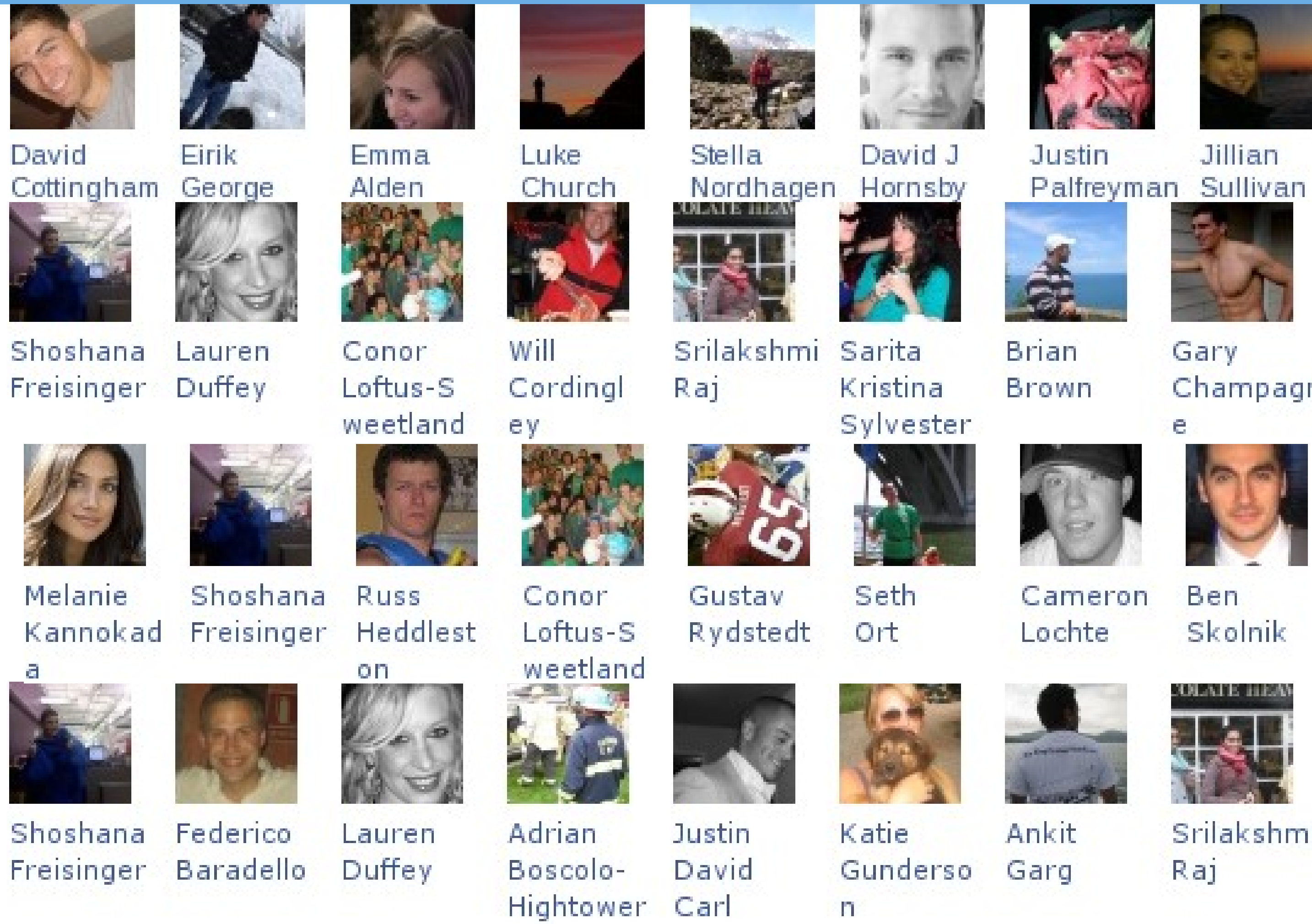
Fun with Tor

UK

Germany

USA

Australia



Attack Scenario

- Spider all public listings
  - Our experiments crawled 250 k users daily
  - Implies ~800 CPU-days to recover all users

Abstraction

- Take a graph $G = <V,E>$
- Randomly select k out-edges from each node
- Result is a sampled graph $G_k = <V,E_k>$
- Try to approximate $f(G) \approx f_{\text{approx}}(G_k)$

Approximable Functions

- Node Degree
- Dominating Set
- Betweenness Centrality
- Path Length
- Community Structure

Experimental Data

- Crawled networks for Stanford, Harvard universities
- Representative sub-networks

	# Users	Mean *d*	Median *d*
Stanford	15043	125	90
Harvard	18273	116	76

Back To Our Abstraction

- Take a graph $G = <V,E>$

- Randomly select k out-edges from each node

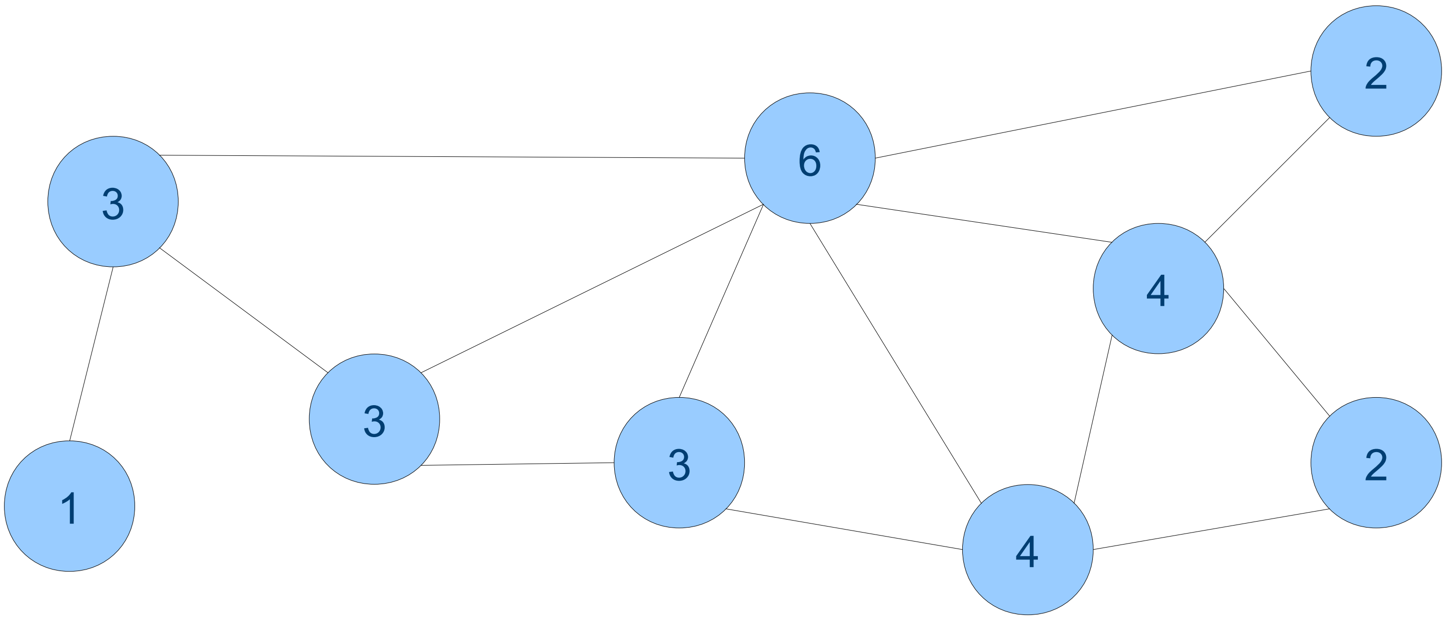
- Result is a sampled graph $G_k = <V,E_k>$

- Try to approximate $f(G) \approx f_{\text{approx}}(G_k)$

Estimating Degrees

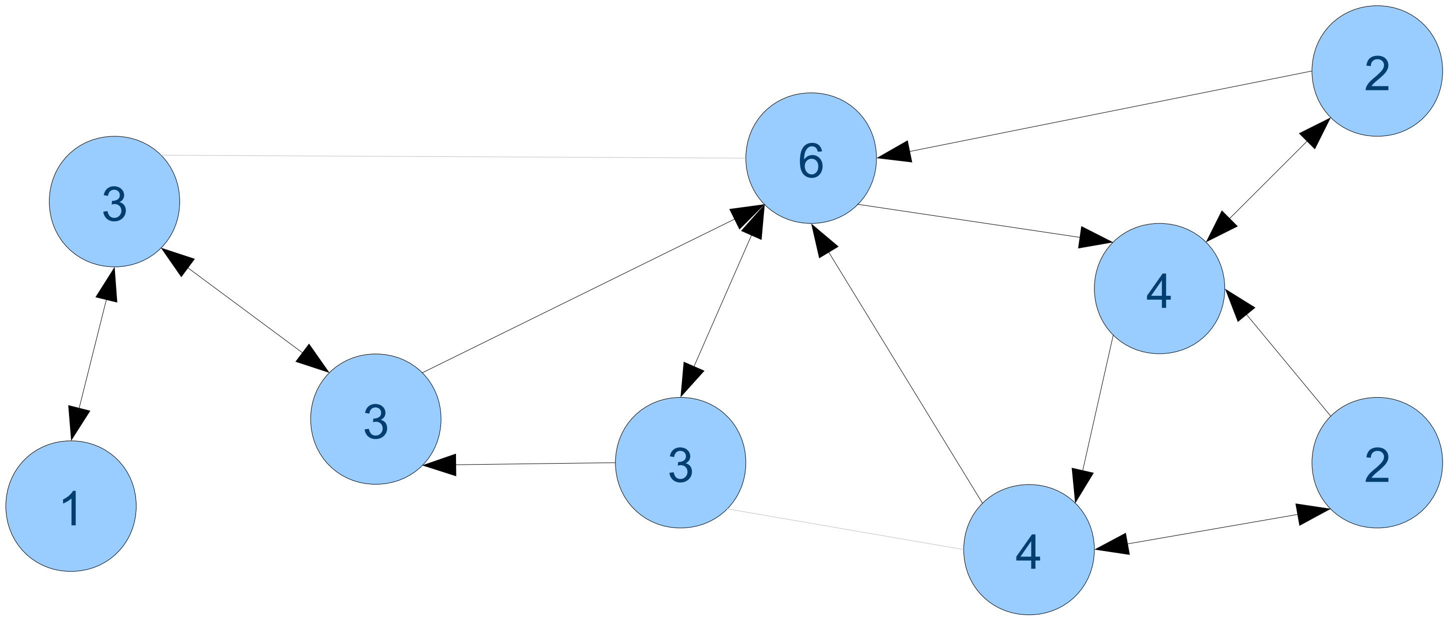
- Convert sampled graph into a directed graph
  - Edges originate at the node where they were seen
- Learn exact degree for nodes with degree < *k*
  - Less than k out-edges
- Get random sample for nodes with degree ≥ *k*
  - Many have more than *k* in-edges

Estimating Degrees



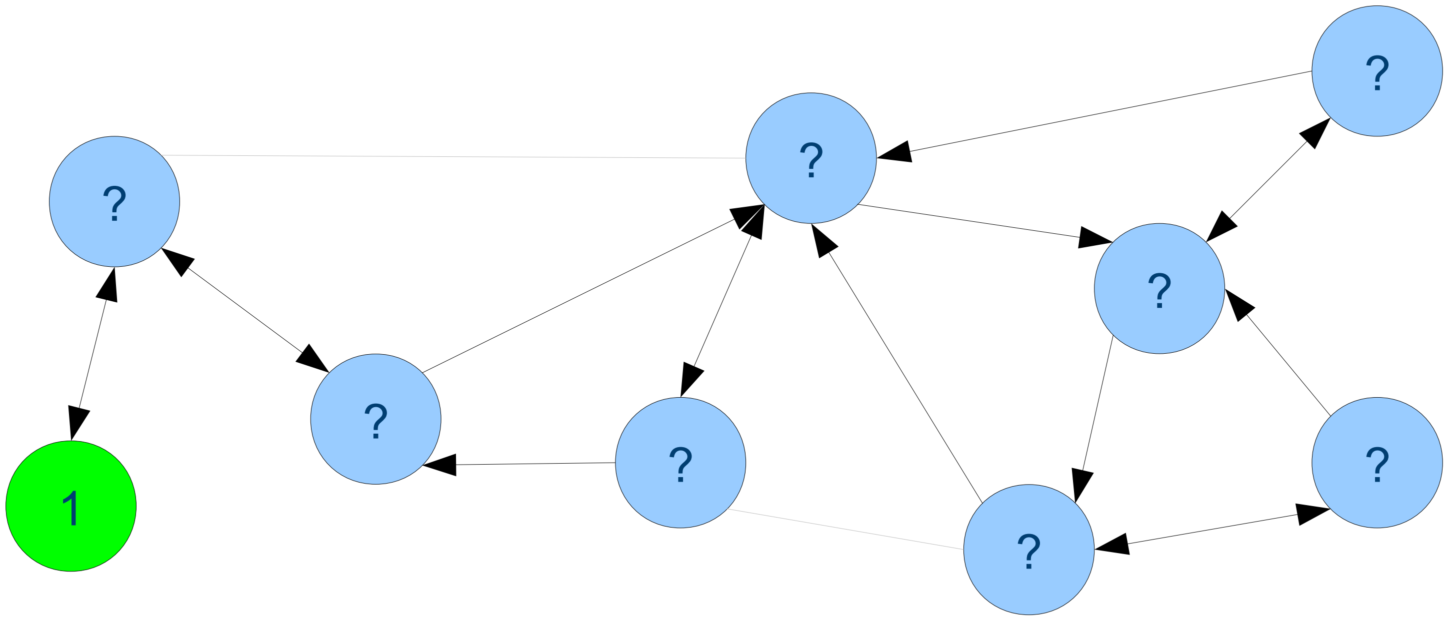
Average Degree: 3.5

Estimating Degrees



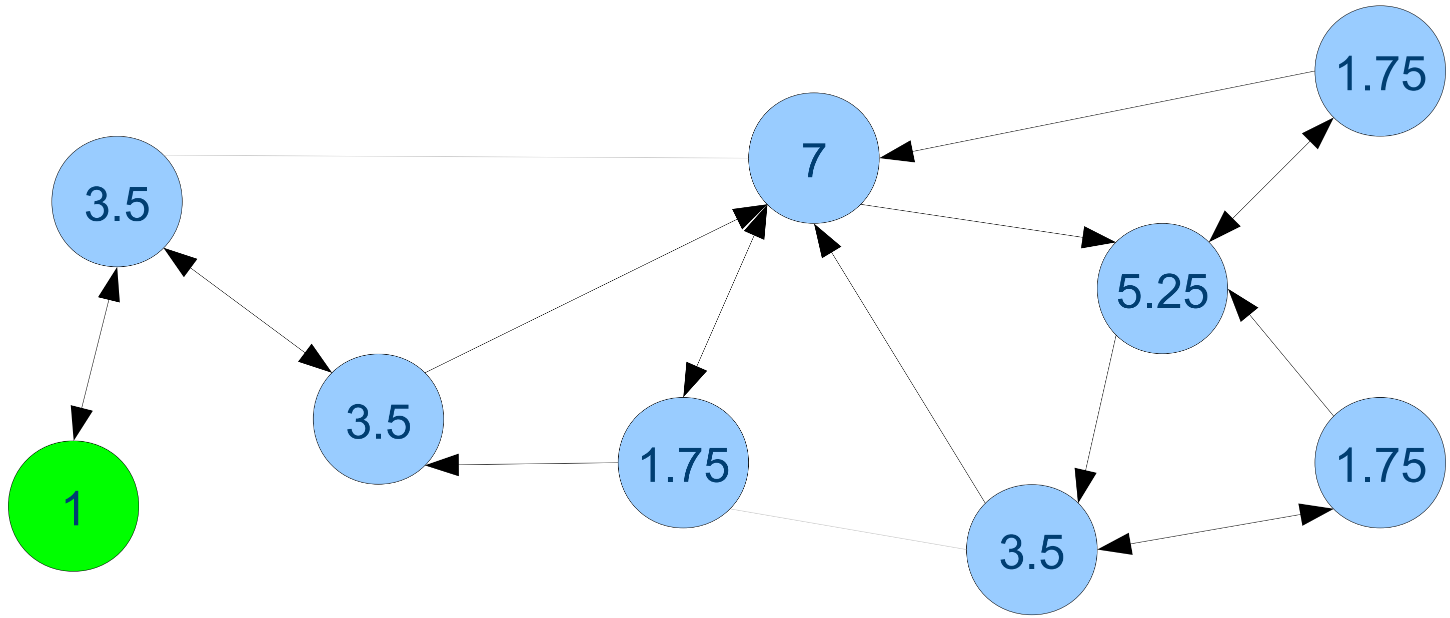
Sampled with k=2

Estimating Degrees



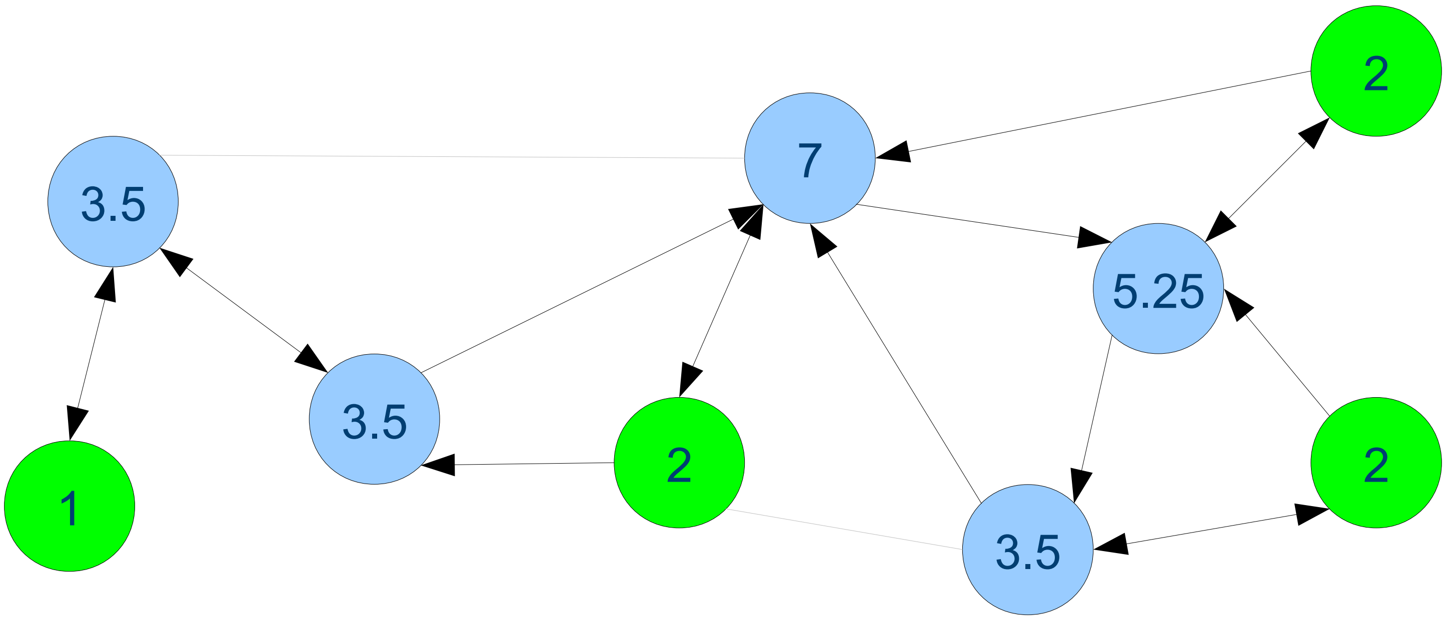
Degree known exactly for one node

Estimating Degrees



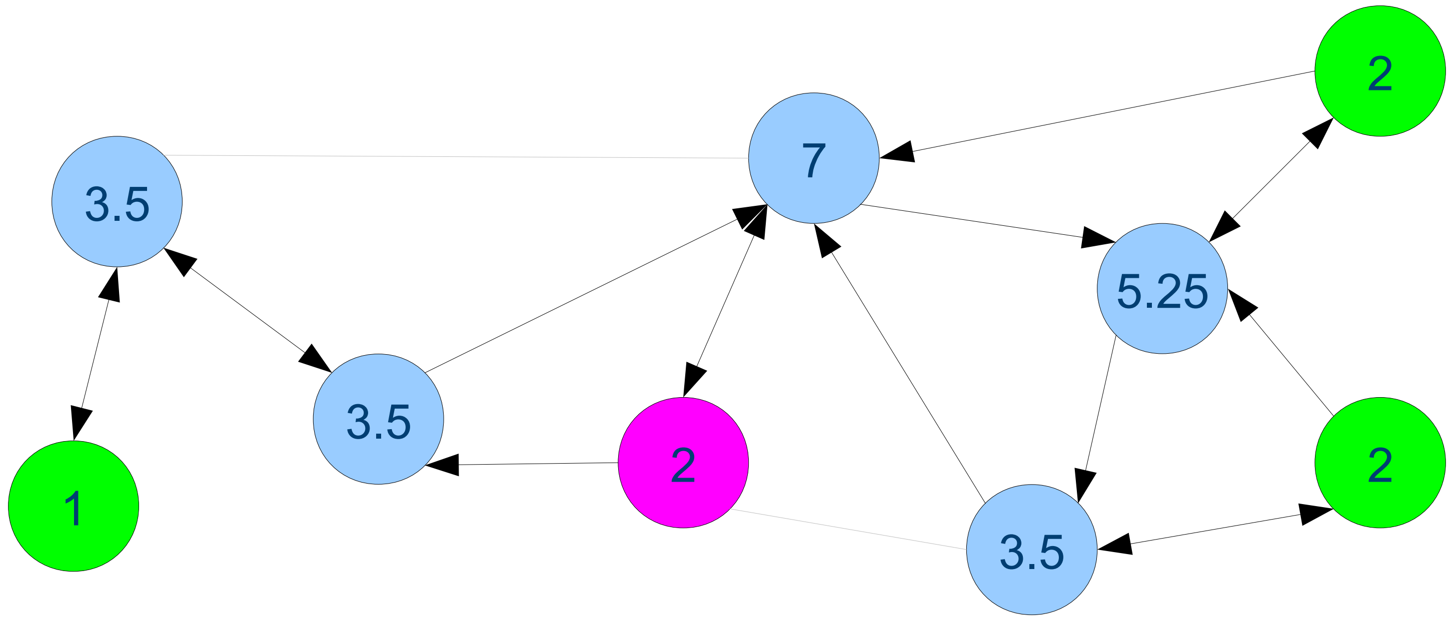
Naïve approach: Multiply in-degree by average degree / *k*

Estimating Degrees



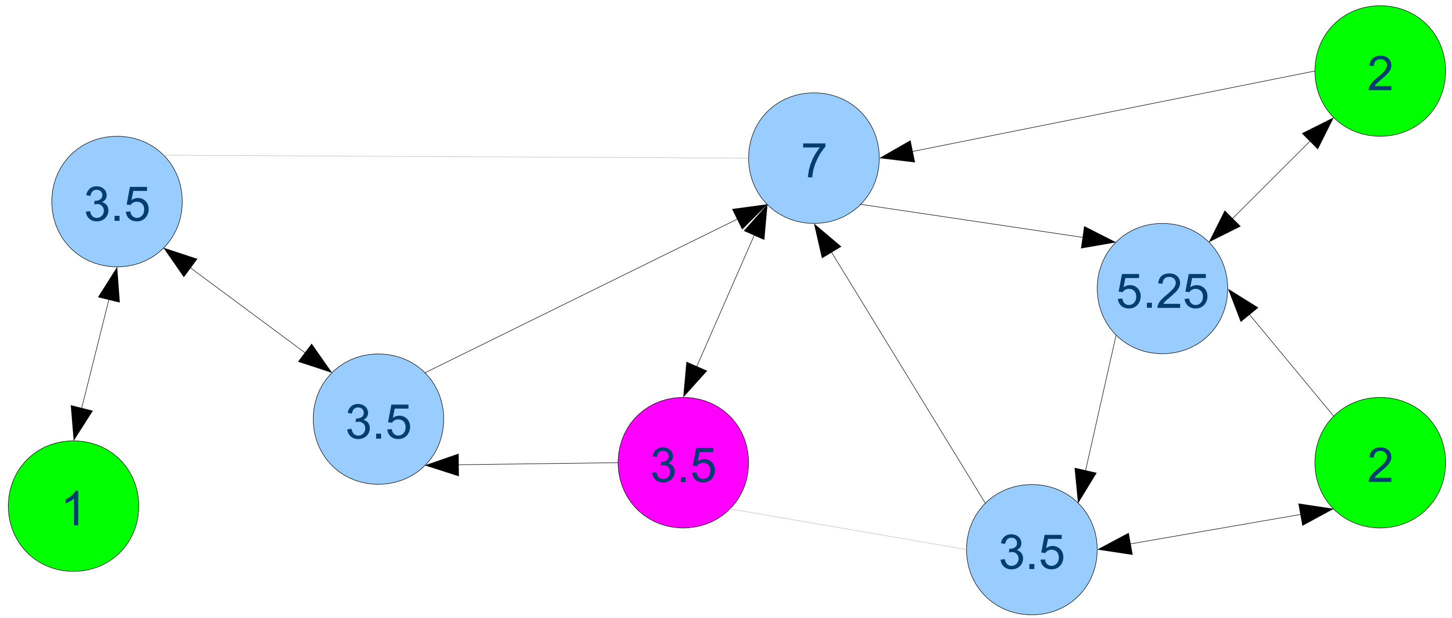
Raise estimates which are less than *k*

Estimating Degrees



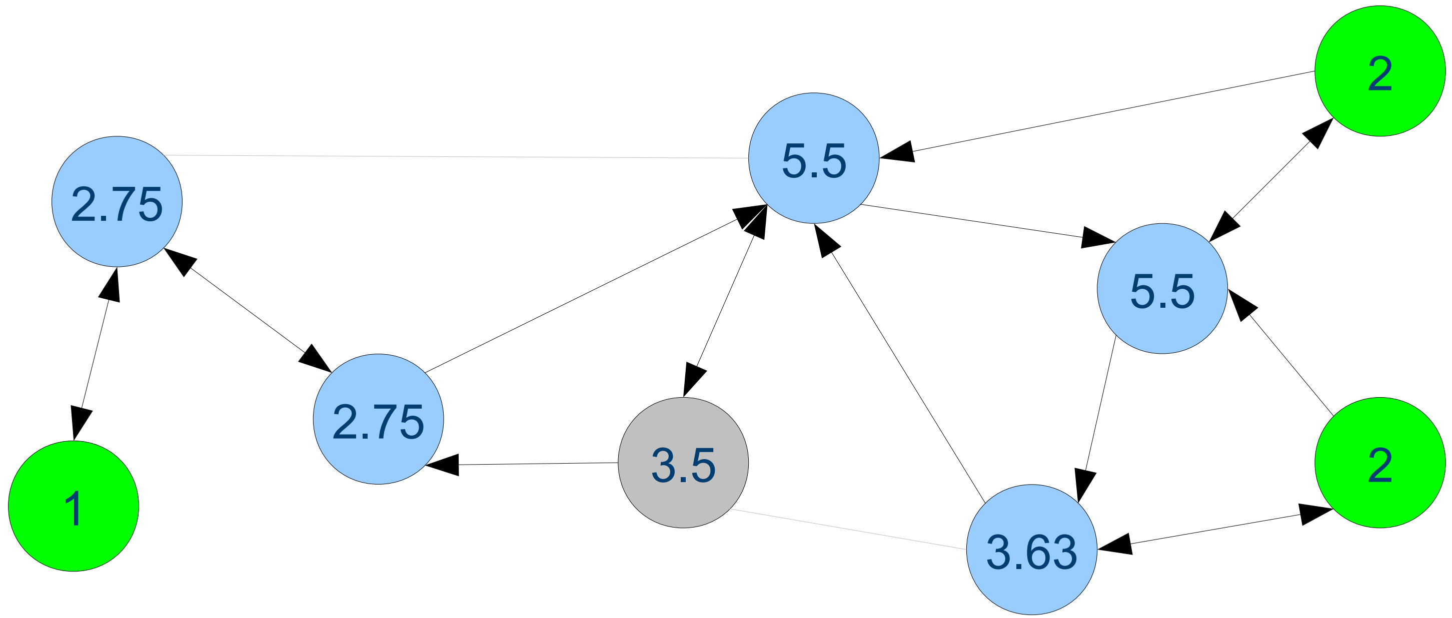
Nodes with high-degree neighbors underestimated

Estimating Degrees



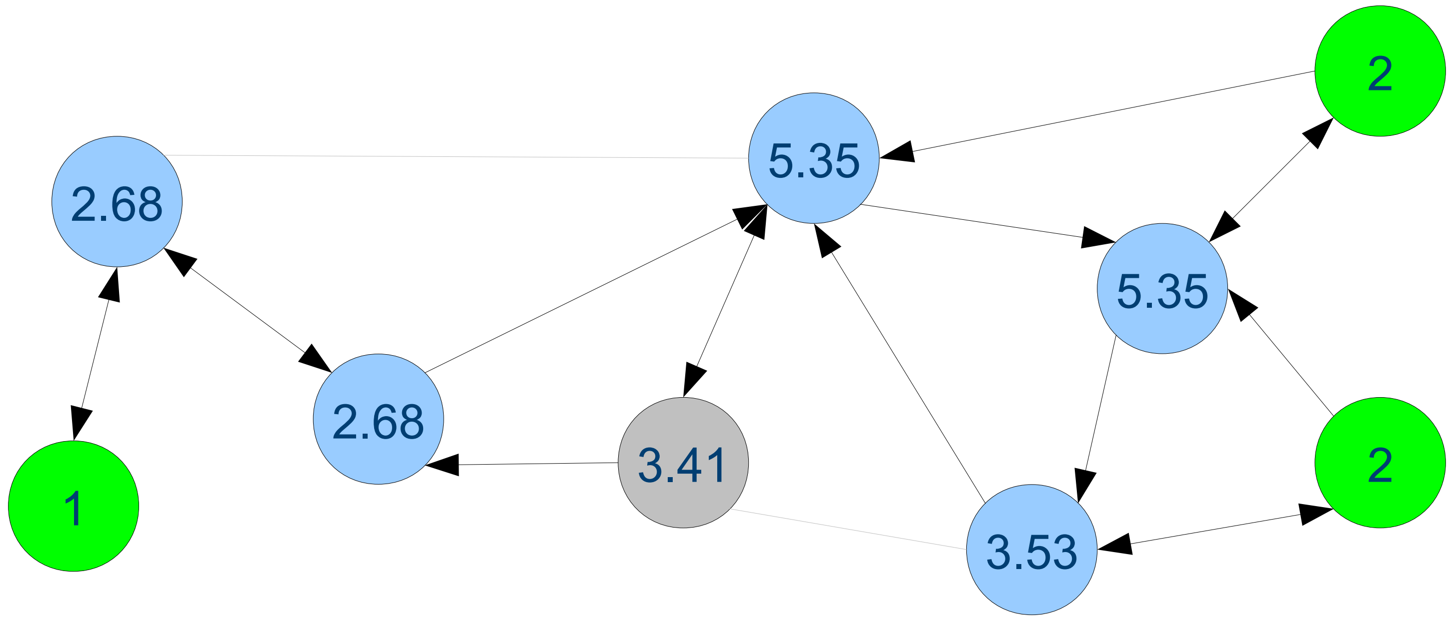
Iteratively scale by current estimate / k in each step

Estimating Degrees



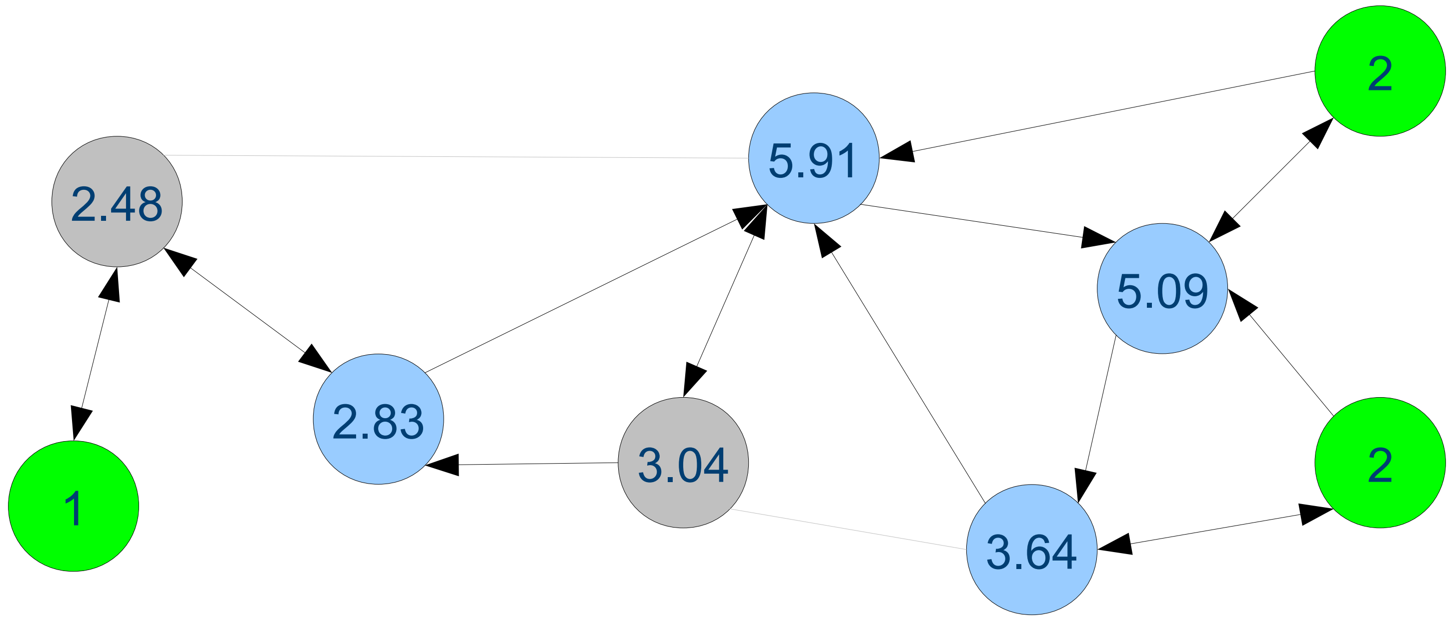
After 1 iteration

Estimating Degrees



Normalise to estimated total degree

Estimating Degrees

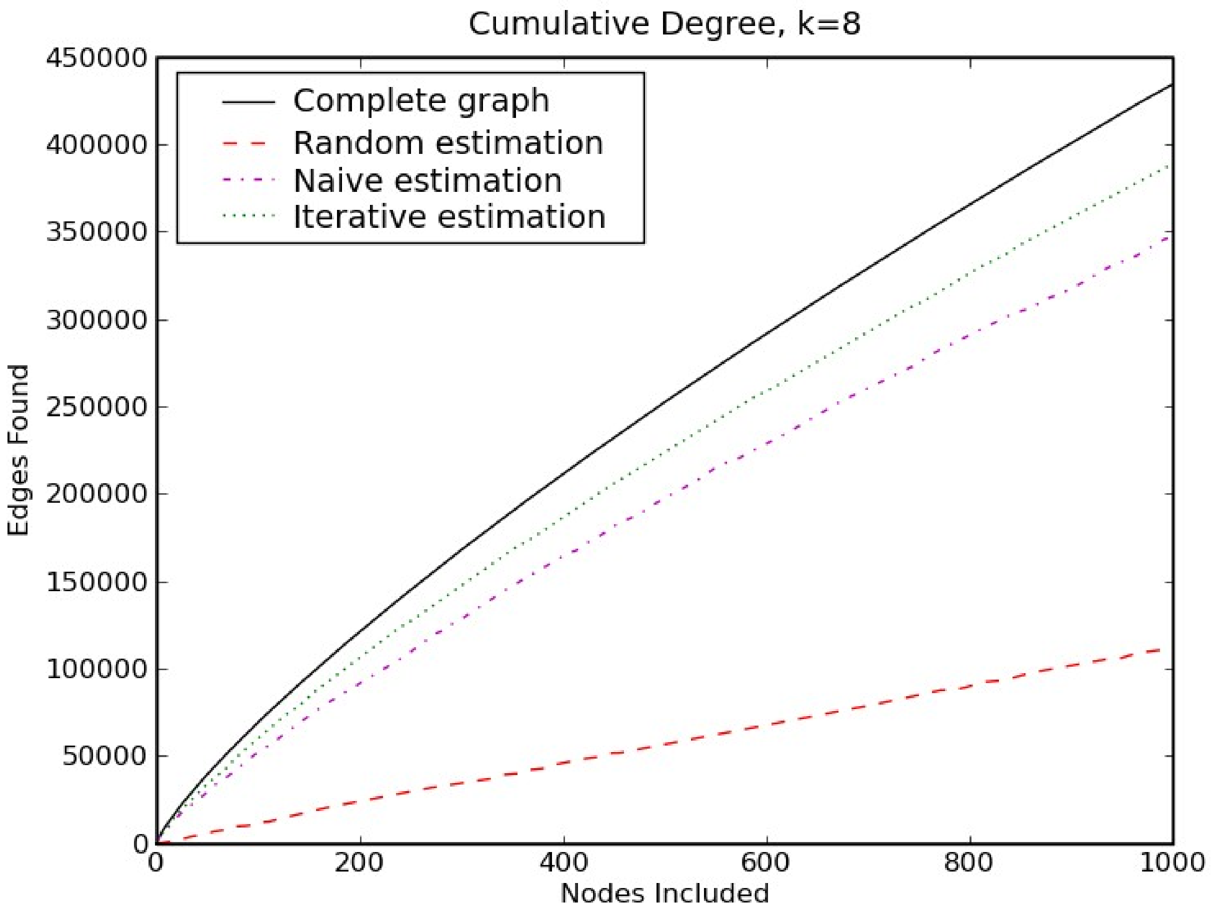


Convergence after n > 10 iterations

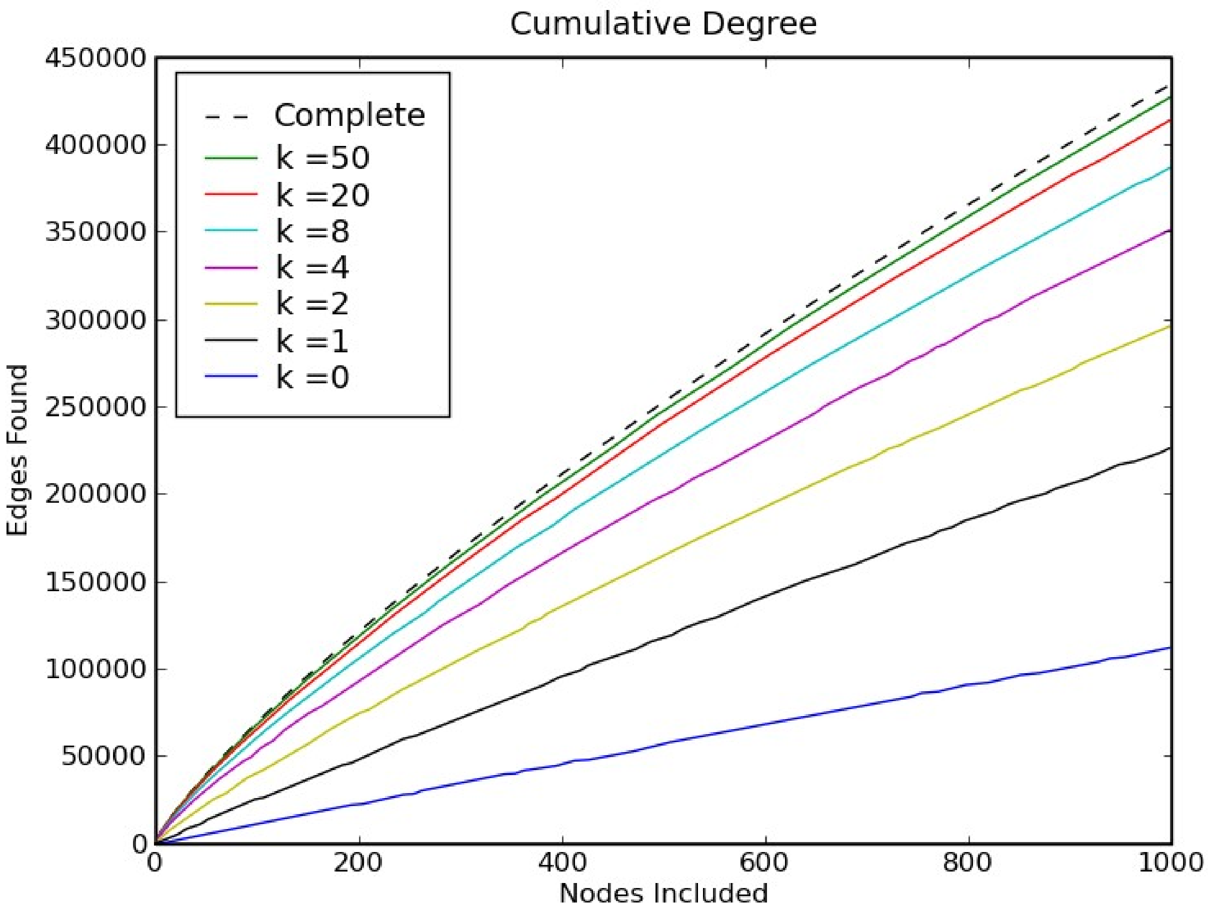
Estimating Degrees

- Converges fast, typically after 10 iterations
- Absolute error is high—38% average
  - Reduced to 23% for nodes with $d \geq 50$
- Still accurately can pick high degree nodes

Aggregate of *x* highest-degree nodes



Comparison of sampling parameters



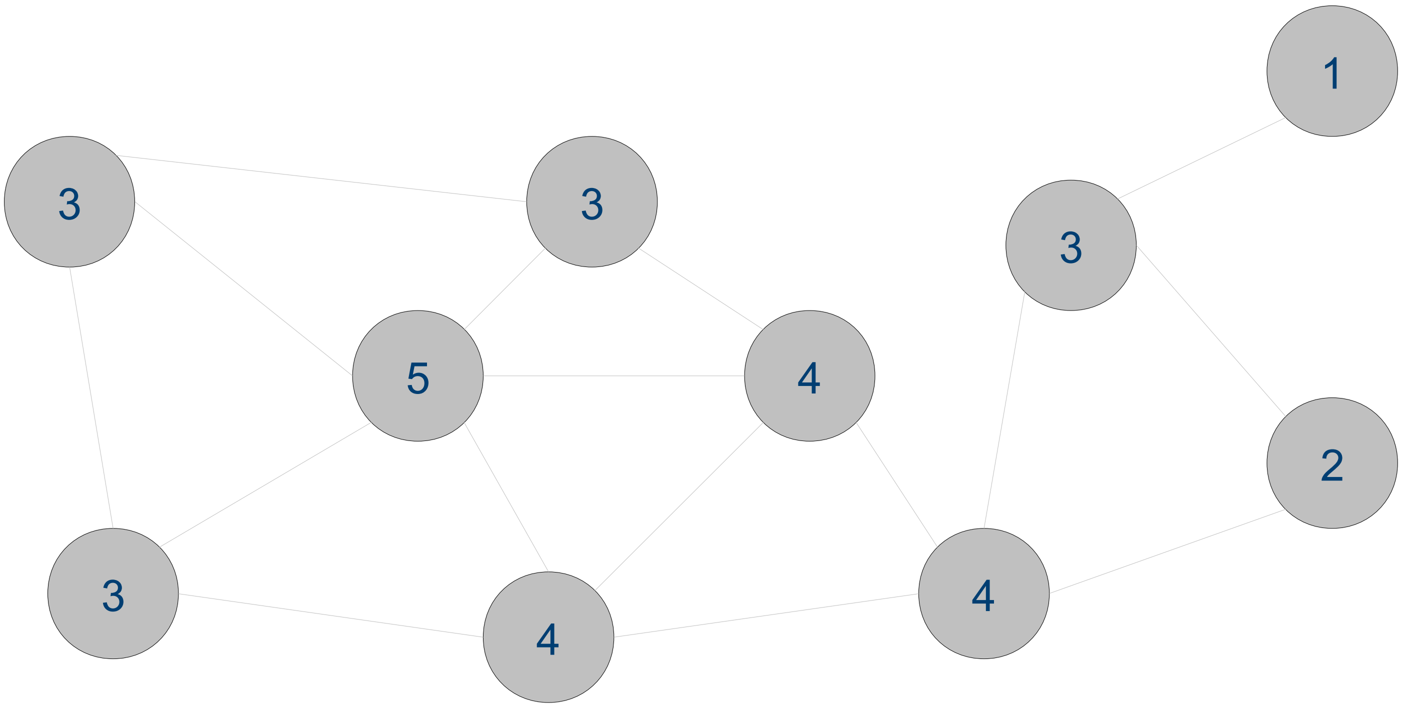
Dominating Sets

- Set of Nodes $D \subseteq V$ such that

$$D \cup \text{Neighbours}(D) = V$$

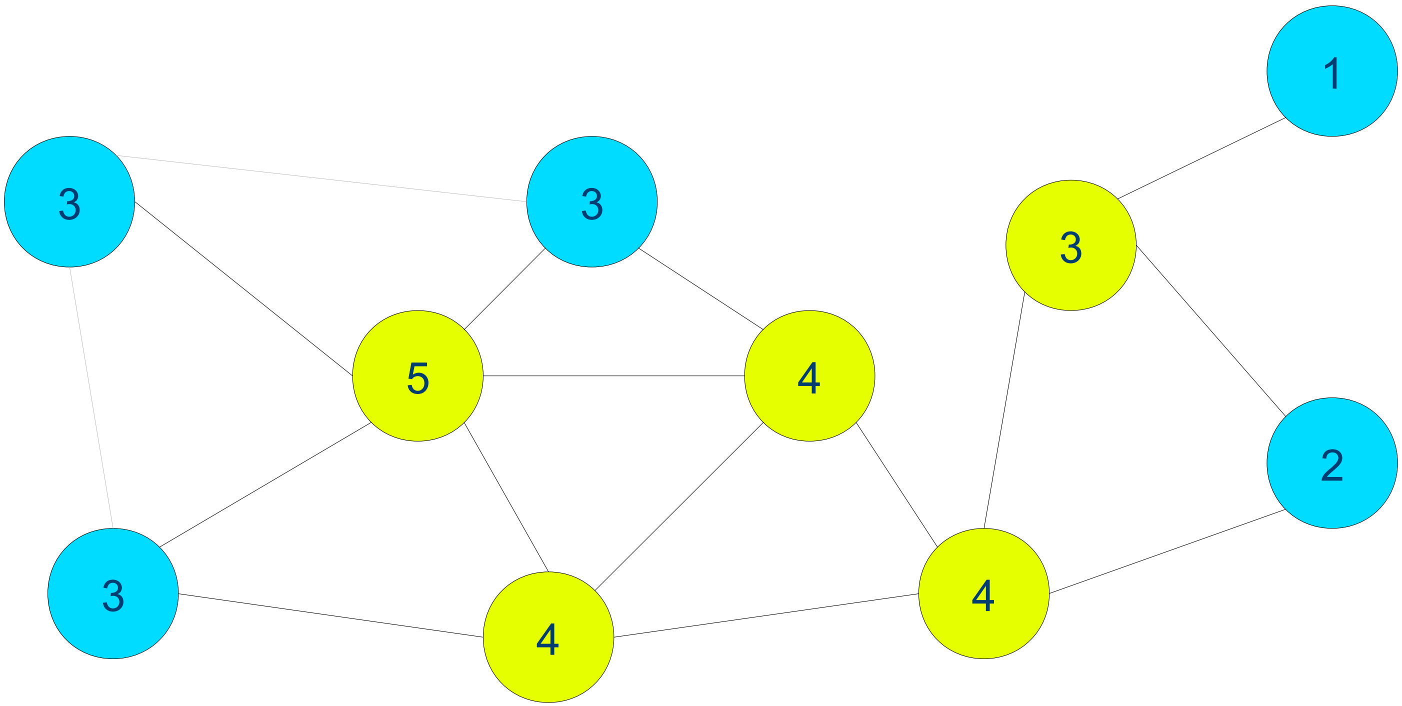
- Set allows viewing the entire network
- Also useful for marketing, trend-setting

Dominating Sets



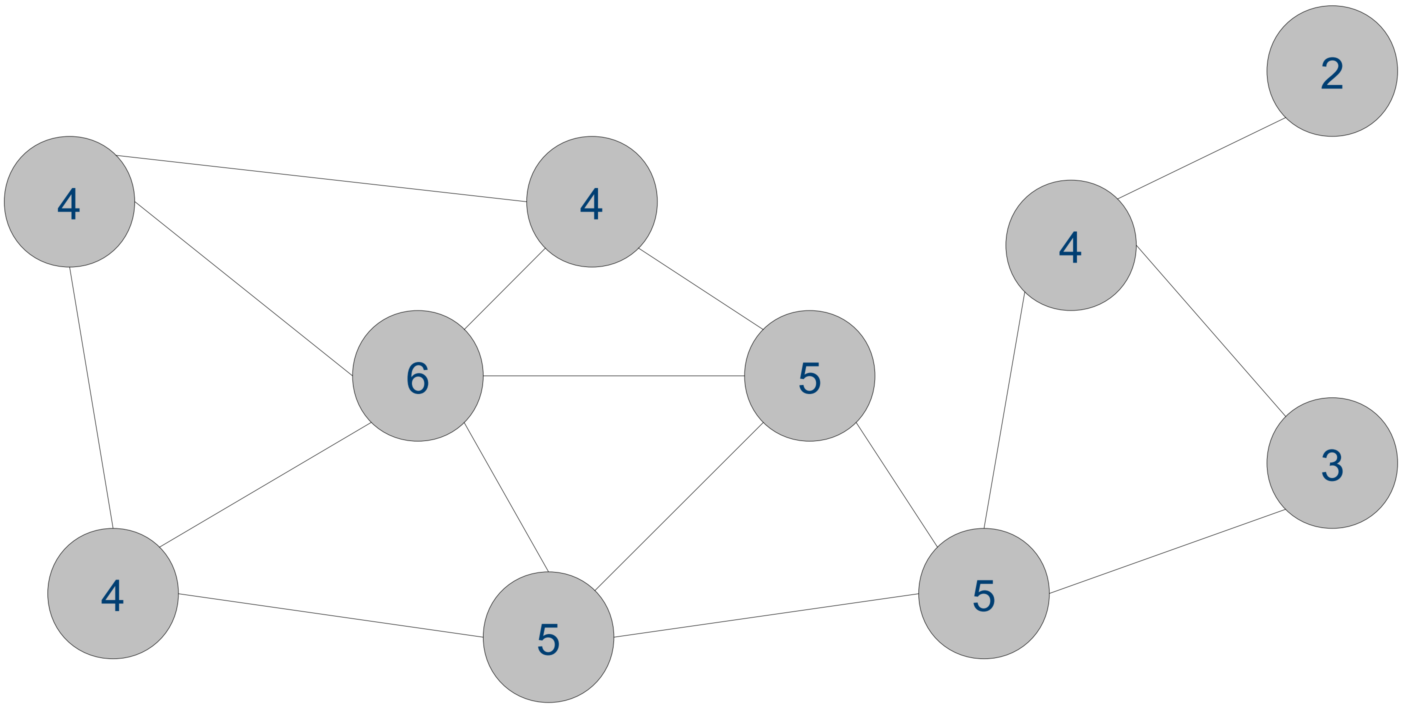
Trivial Algorithm: Select High-Degree Nodes in Order

Dominating Sets



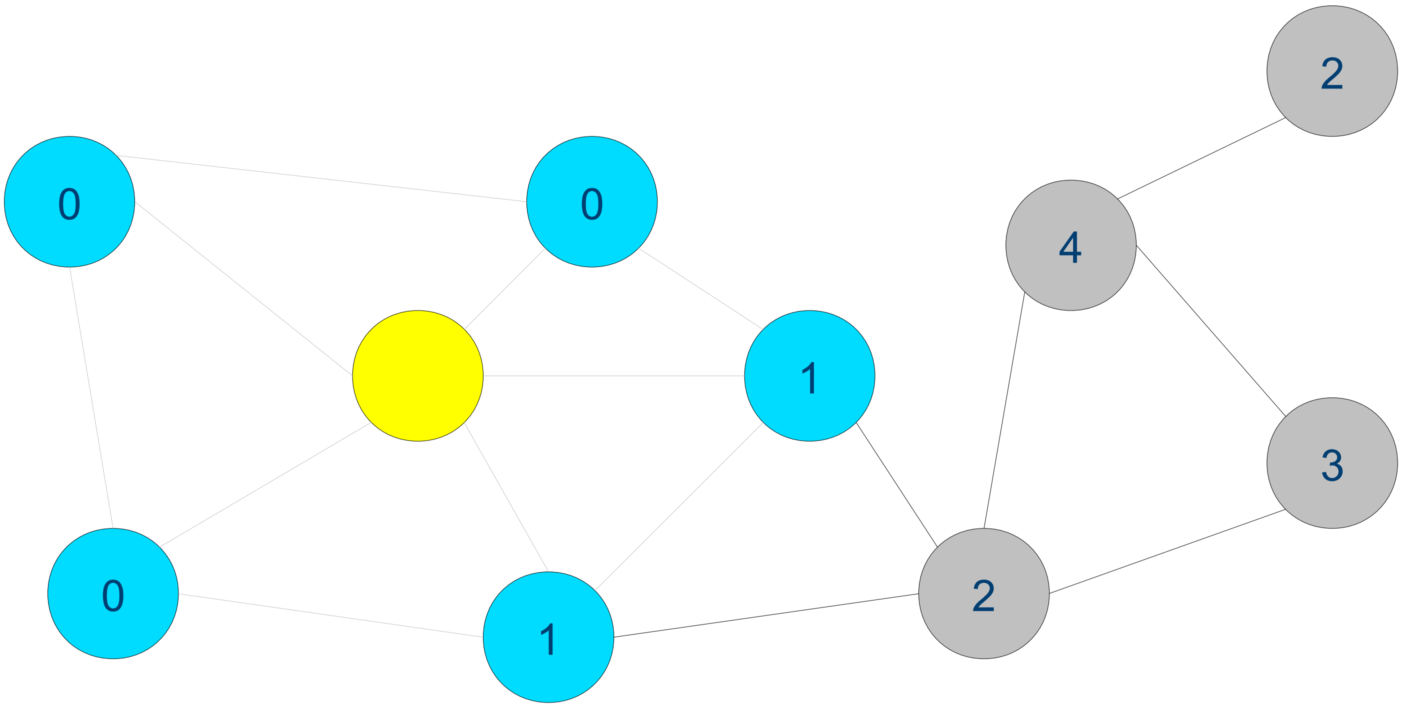
In fact, finding minimal dominating set is NP-complete

Dominating Sets



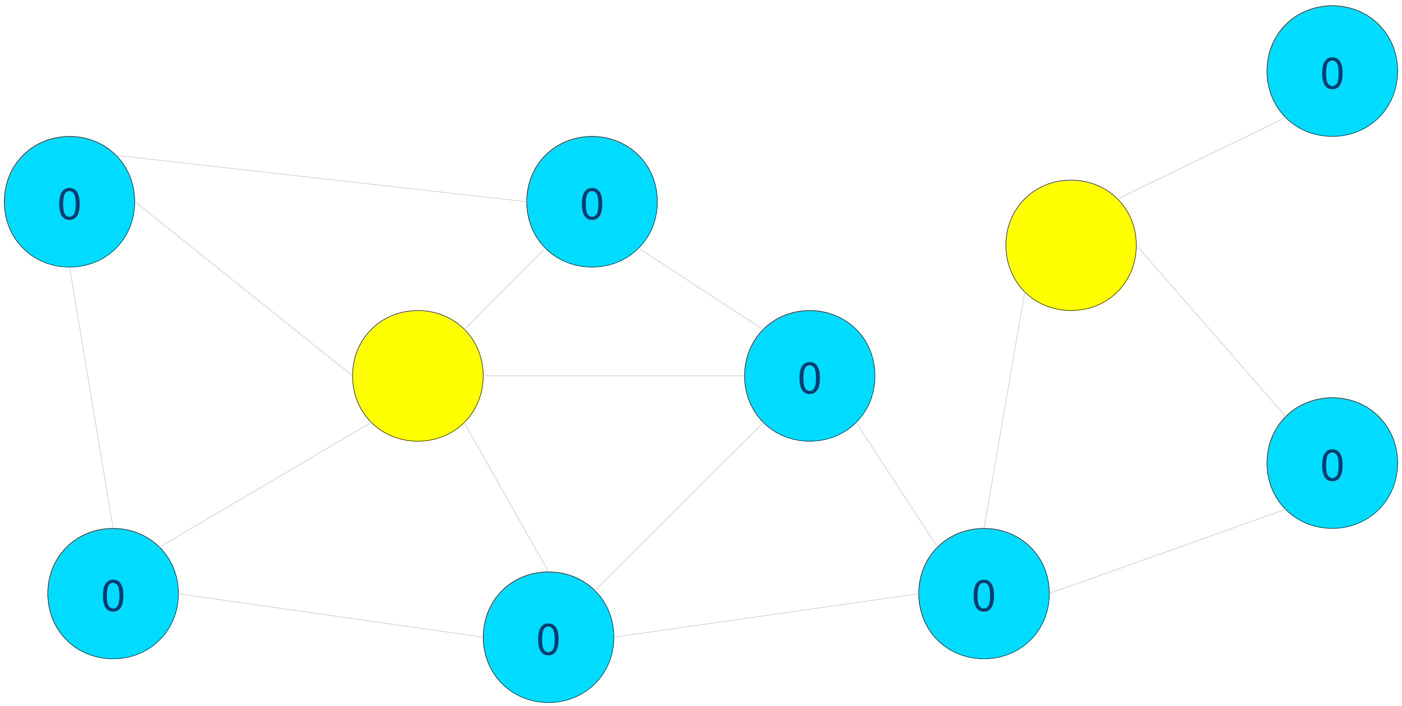
Greedy Algorithm: select for maximal coverage

Dominating Sets



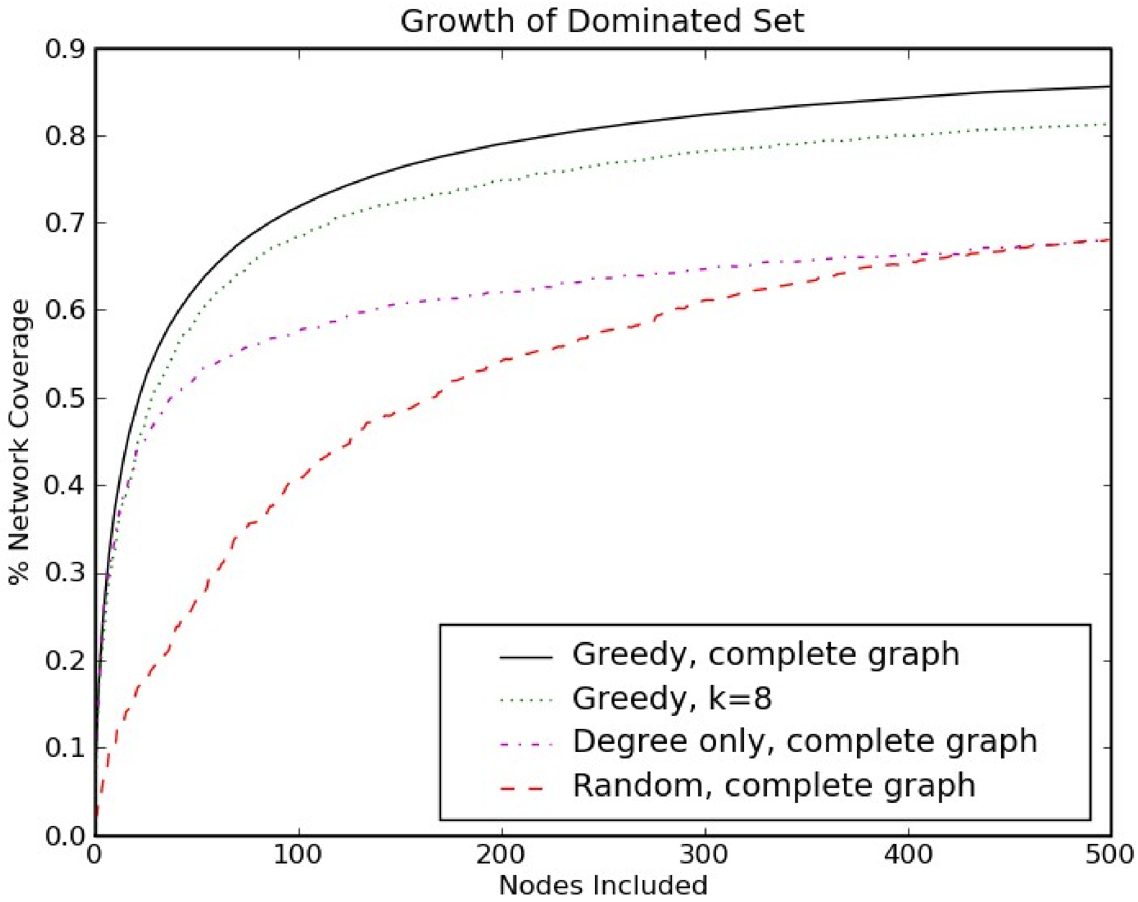
Greedy Algorithm: select for maximal coverage

Dominating Sets

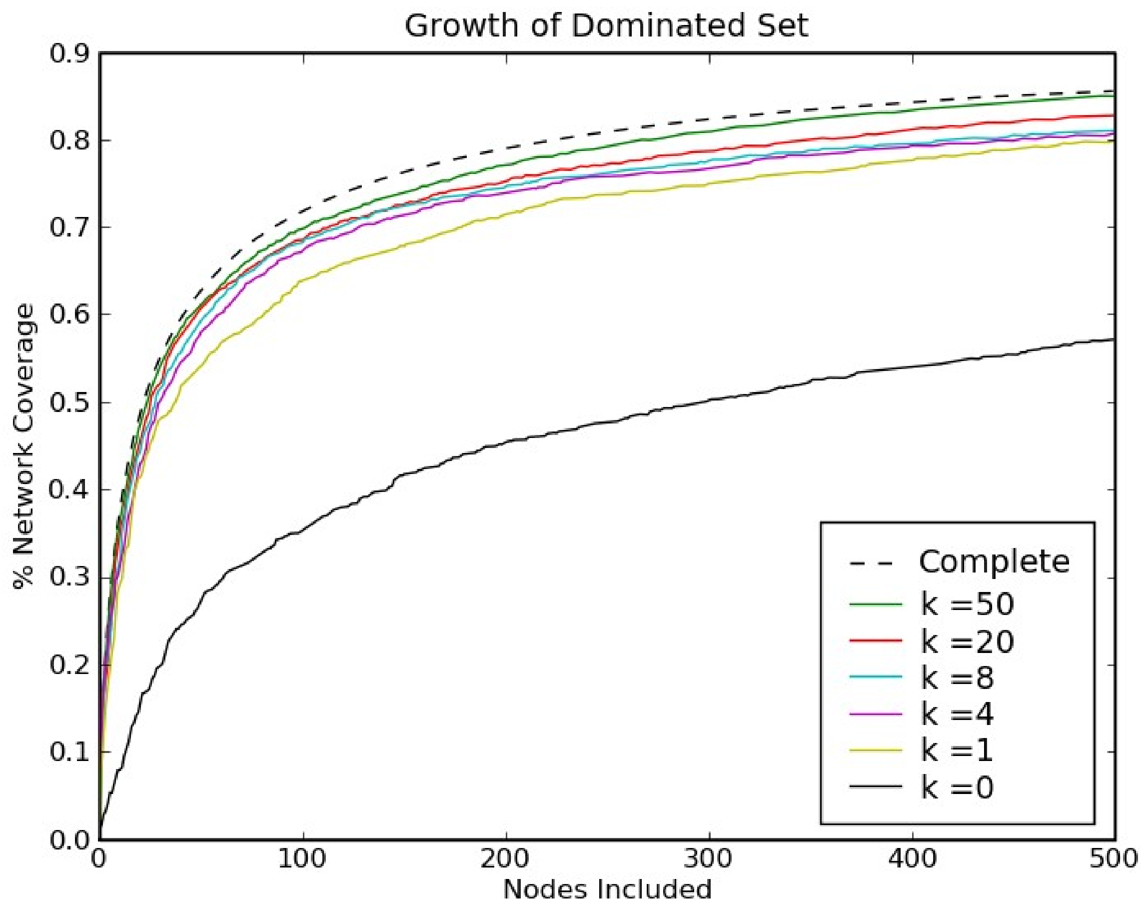


Shown to perform adequately in practice

Works Well on Sampled Graph



Insensitive to Sampling Parameter!



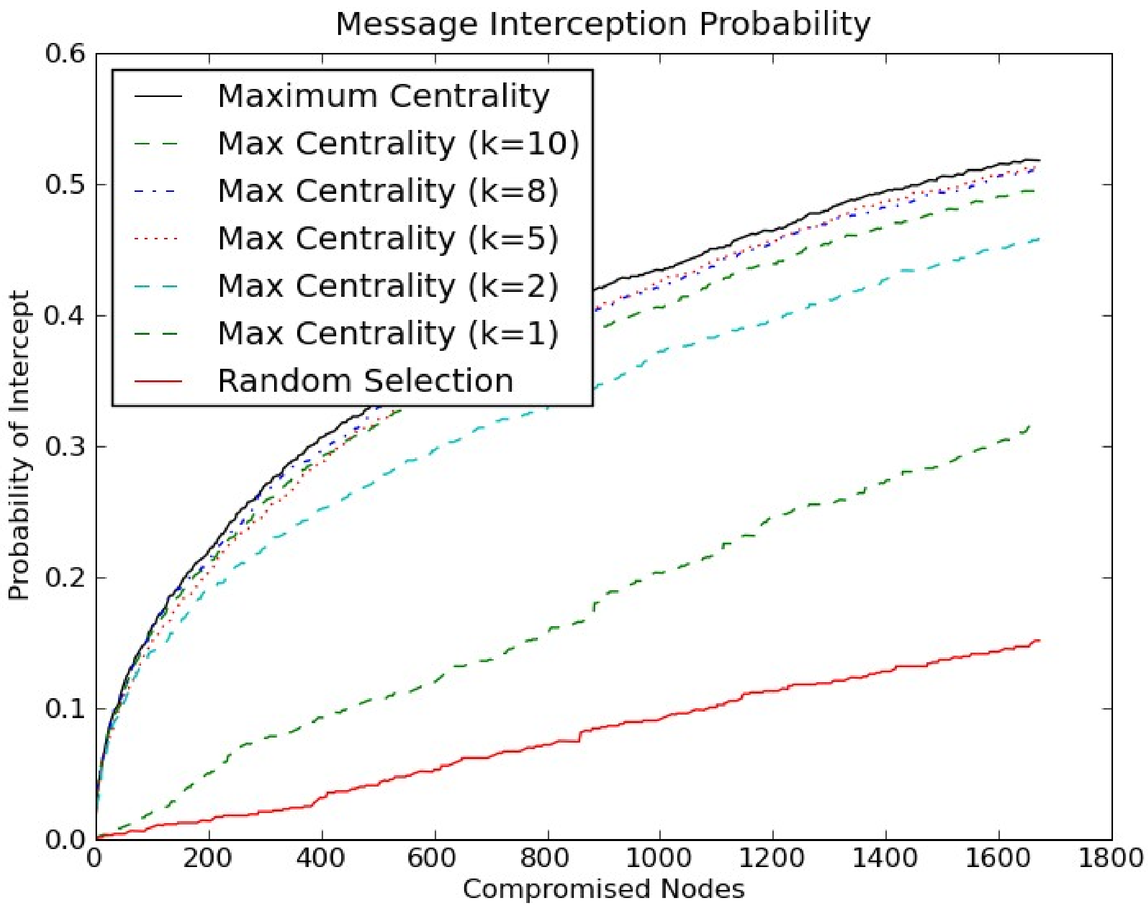
Centrality

- A measure of a node's importance
- *Betweenness centrality*:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

- Measures the shortest paths in the graph that a particular vertex is part of

Centrality



Message Interception Probability
Maximum Centrality
Max Centrality (k=10)
Max Centrality (k=8)
Max Centrality (k=5)
Max Centrality (k=2)
Max Centrality (k=1)
Random Selection
Probability of Intercept
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
200
400
600
800
1000
1200
1400
1600
1800
Compromised Nodes

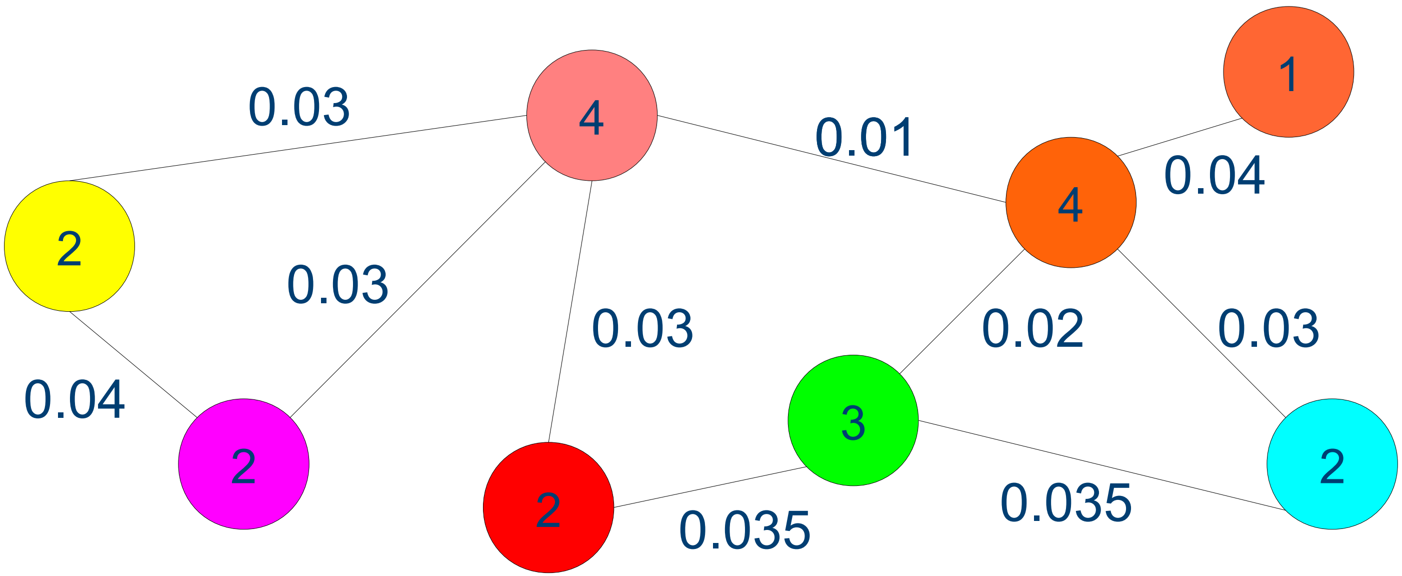
Community Detection

- Goal: Find highly-connected sub-groups
- Measure success by high *modularity*:

$$Q = \frac{1}{2m} \sum_{v,w} \left[A_{vw} - \frac{d(v)d(w)}{2m} \right]$$

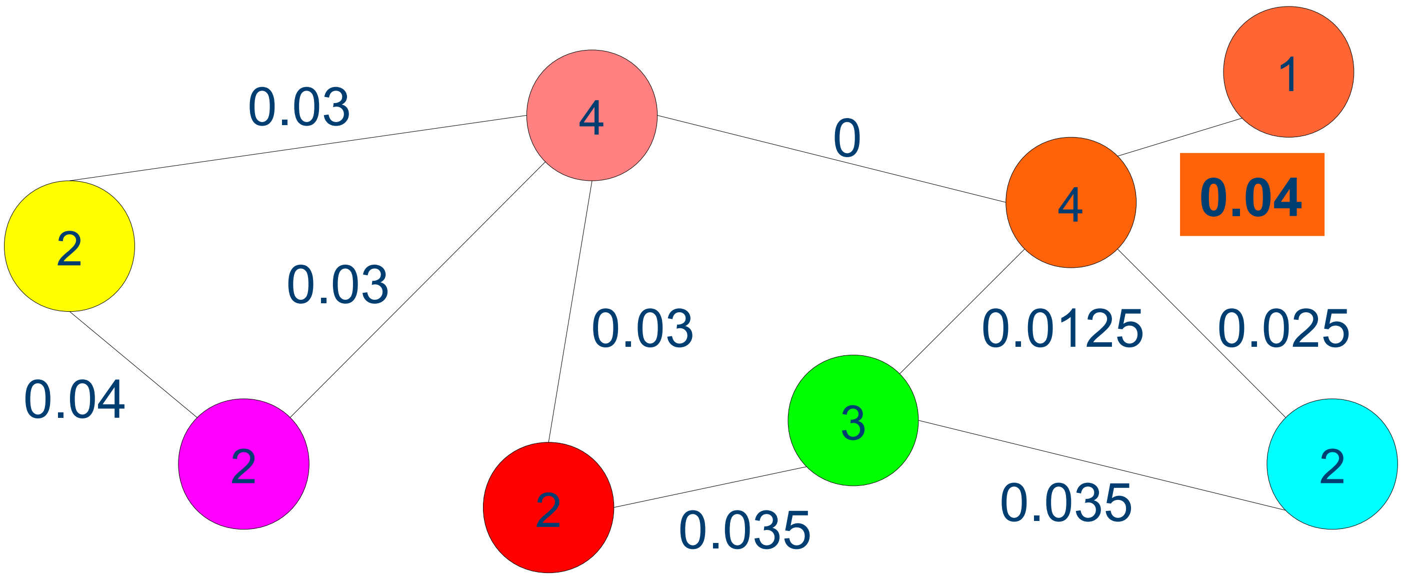
- Ratio of intra-community edges to random
- Normalised to be between -1 and 1

Community Detection



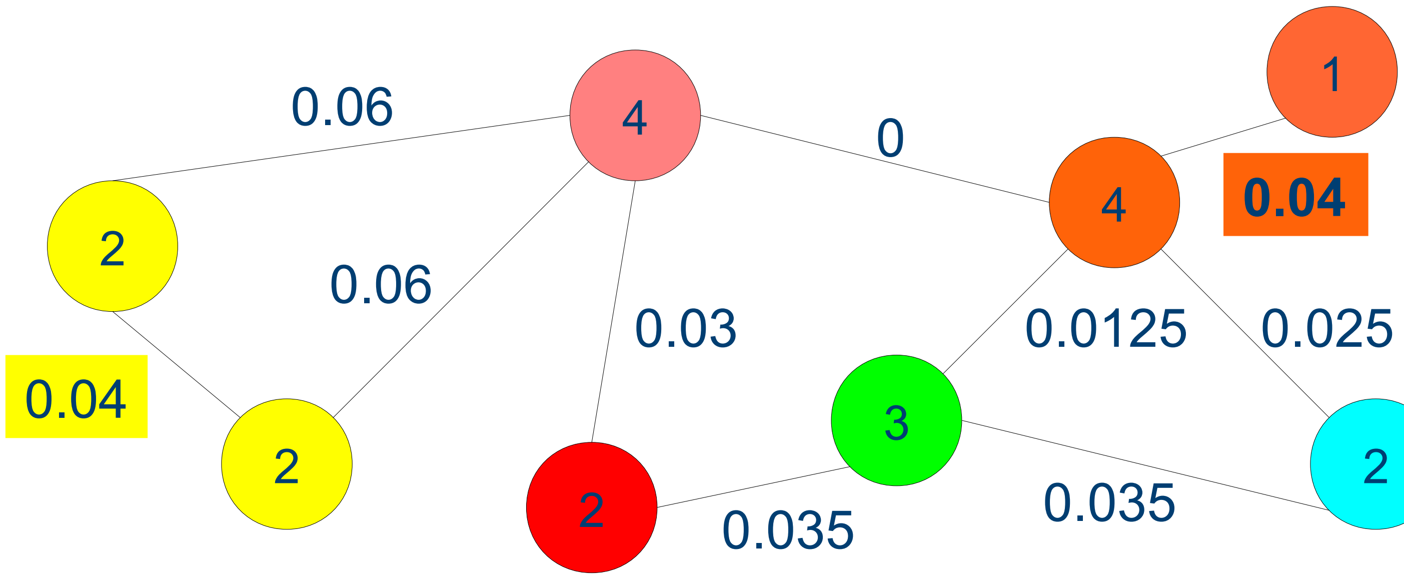
- Clausen et. al 2004 – find maximal modularity in O($n\lg^2 n$)
- Track marginal modularity, update neighbours on each merge

Community Detection



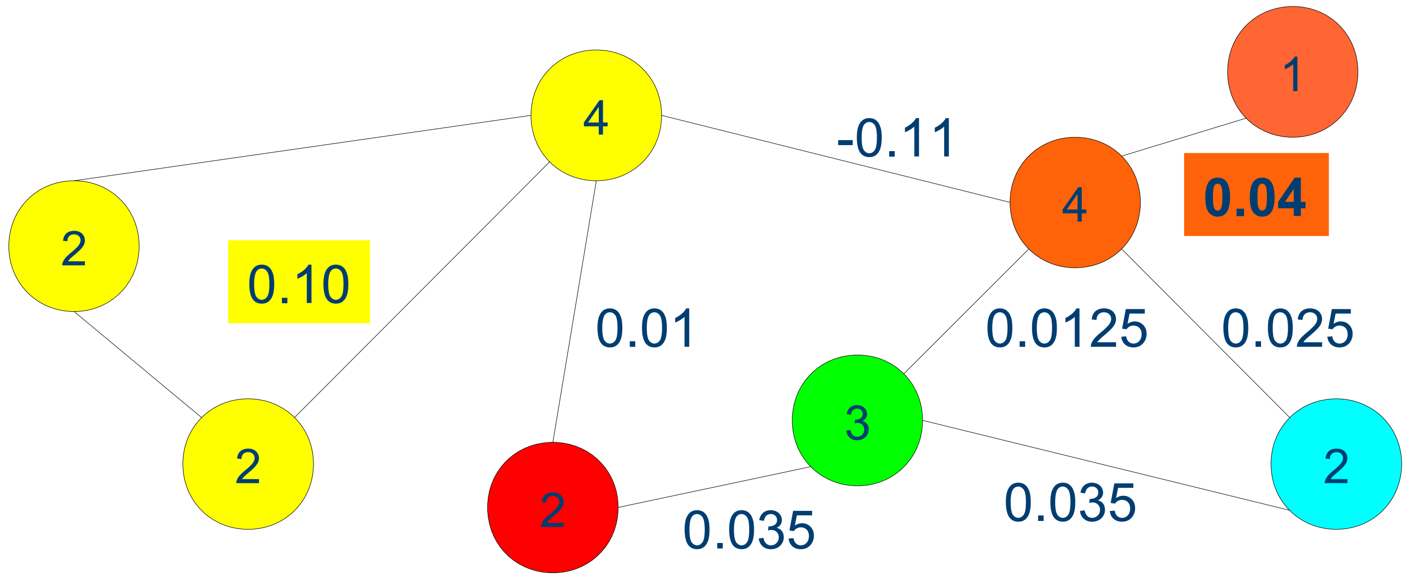
Q=0.04

Community Detection



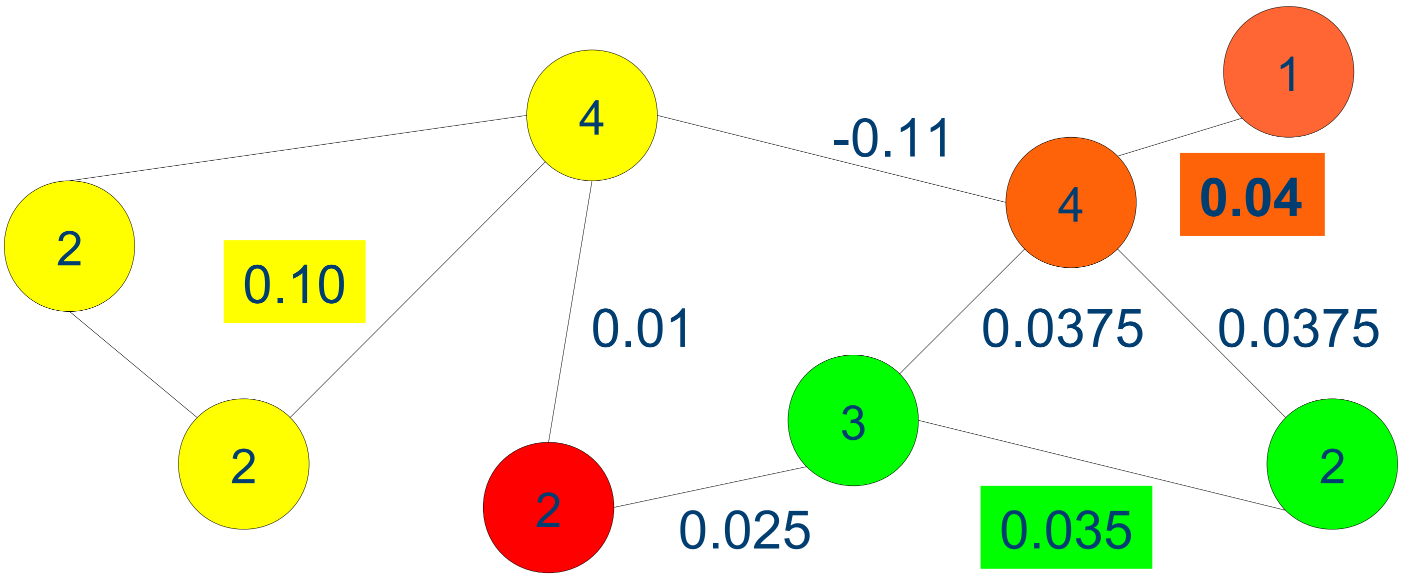
Q=0.08

Community Detection



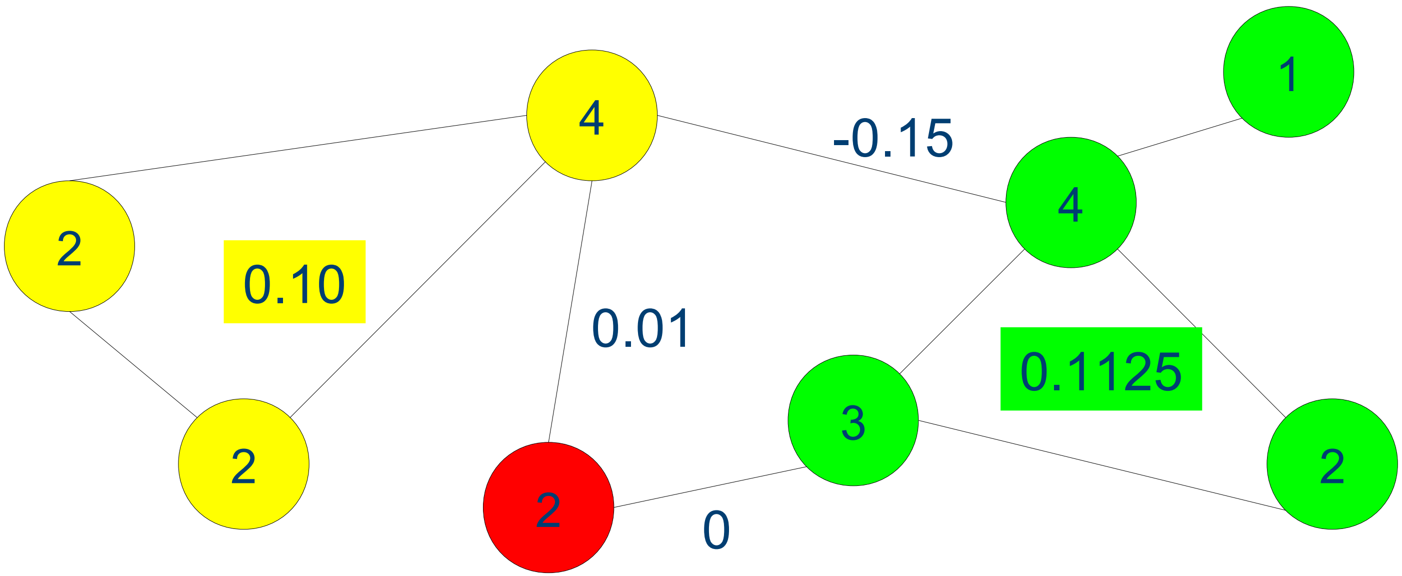
Q=0.14

Community Detection



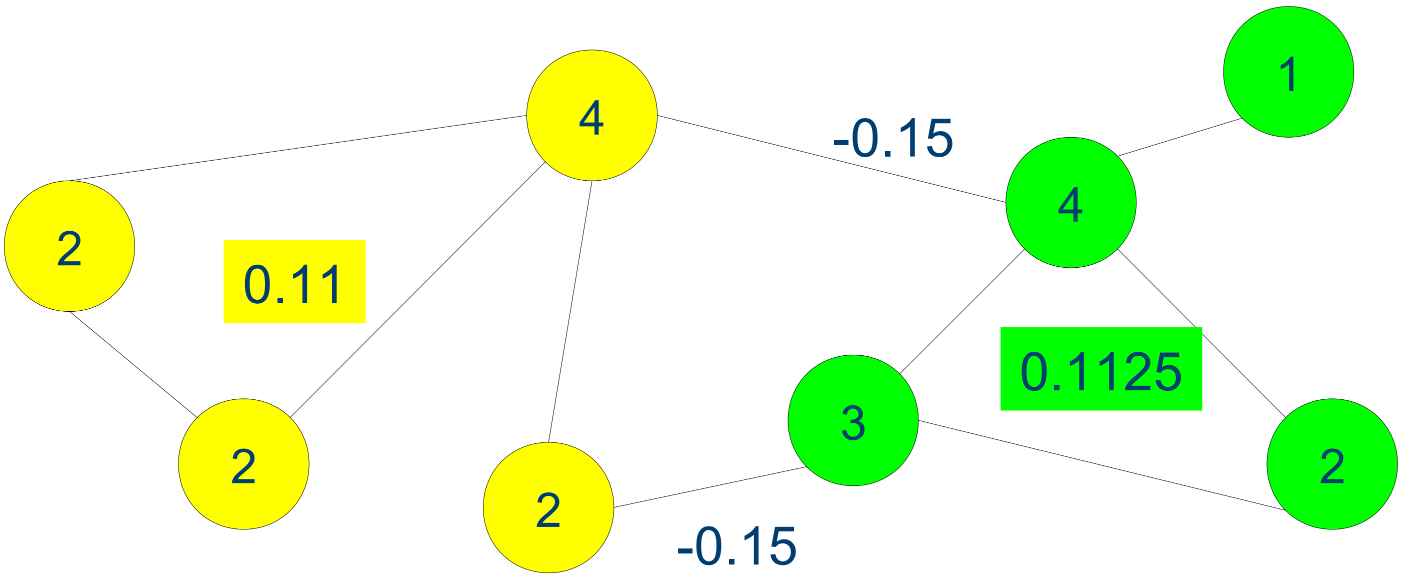
Q=0.175

Community Detection



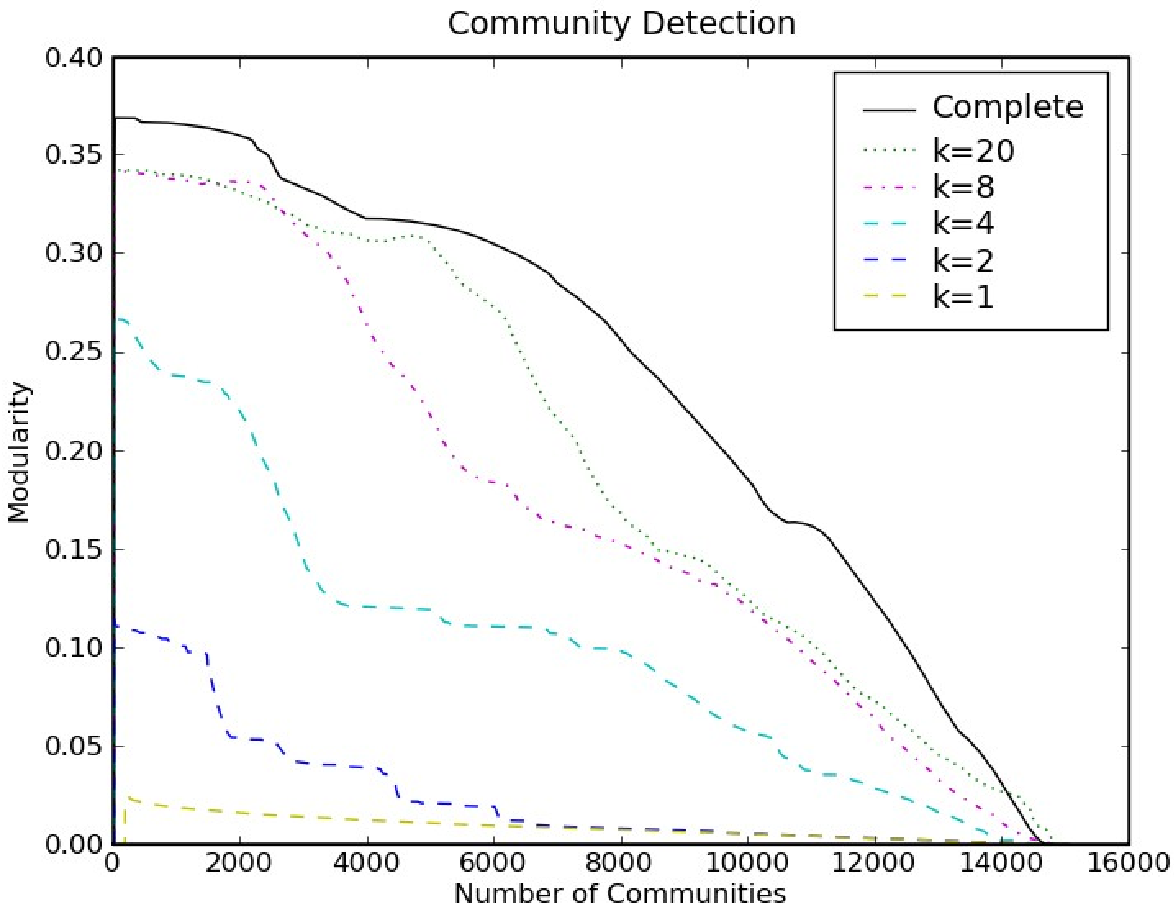
Q=0.2125

Community Detection



Q=0.2225

Community Detection



Conclusions

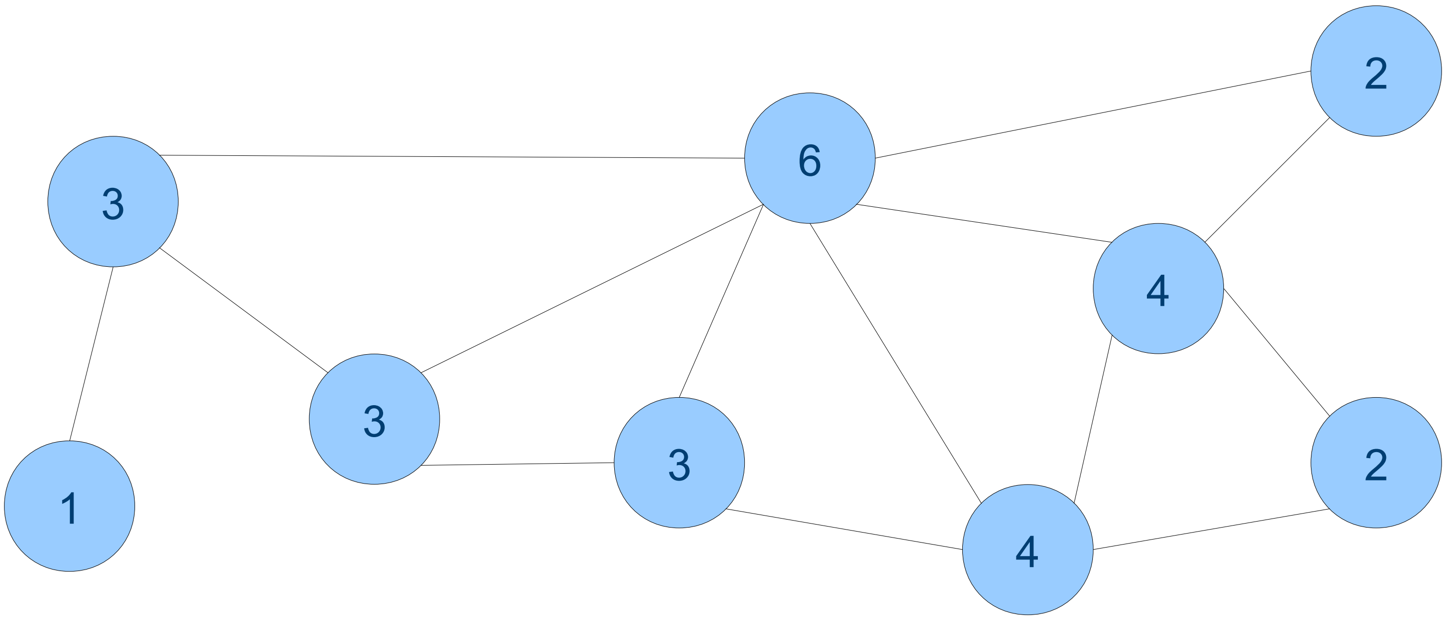
- *k*-sampling of each edge gives away a lot

Conclusions

- *k*-sampling of each edge gives away a lot

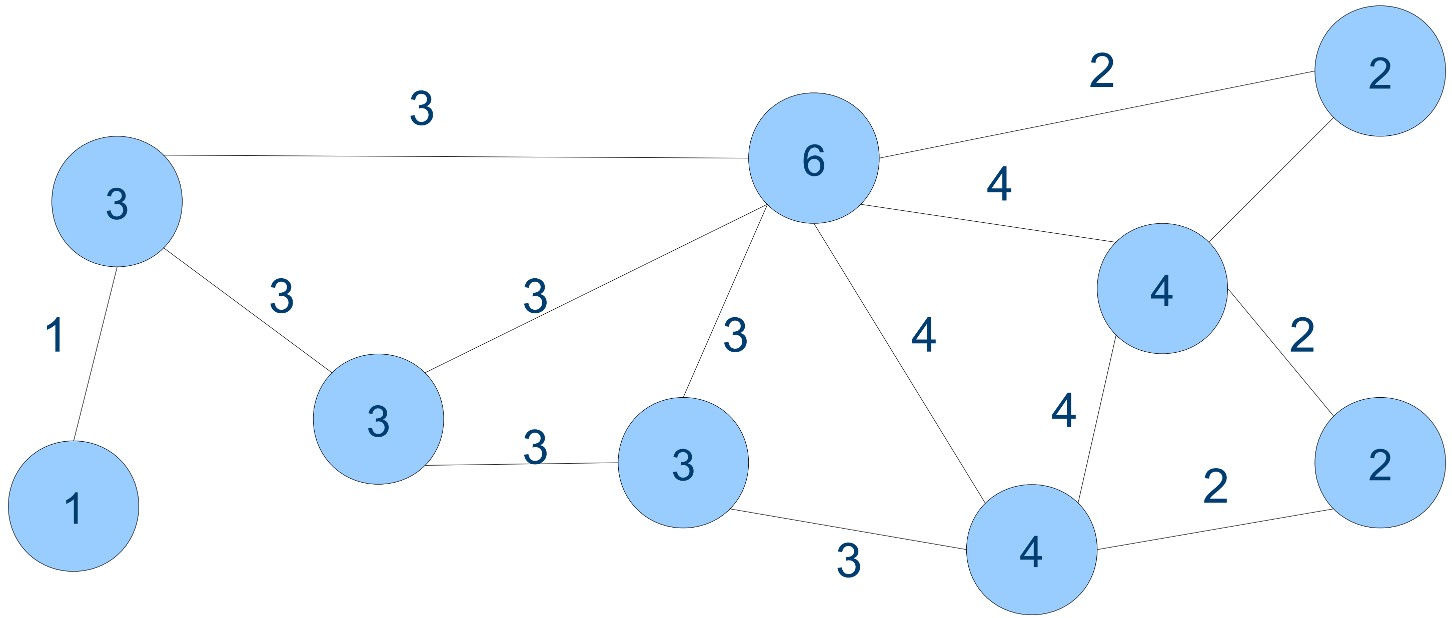
Can we fix it?

Regular subgraph extraction



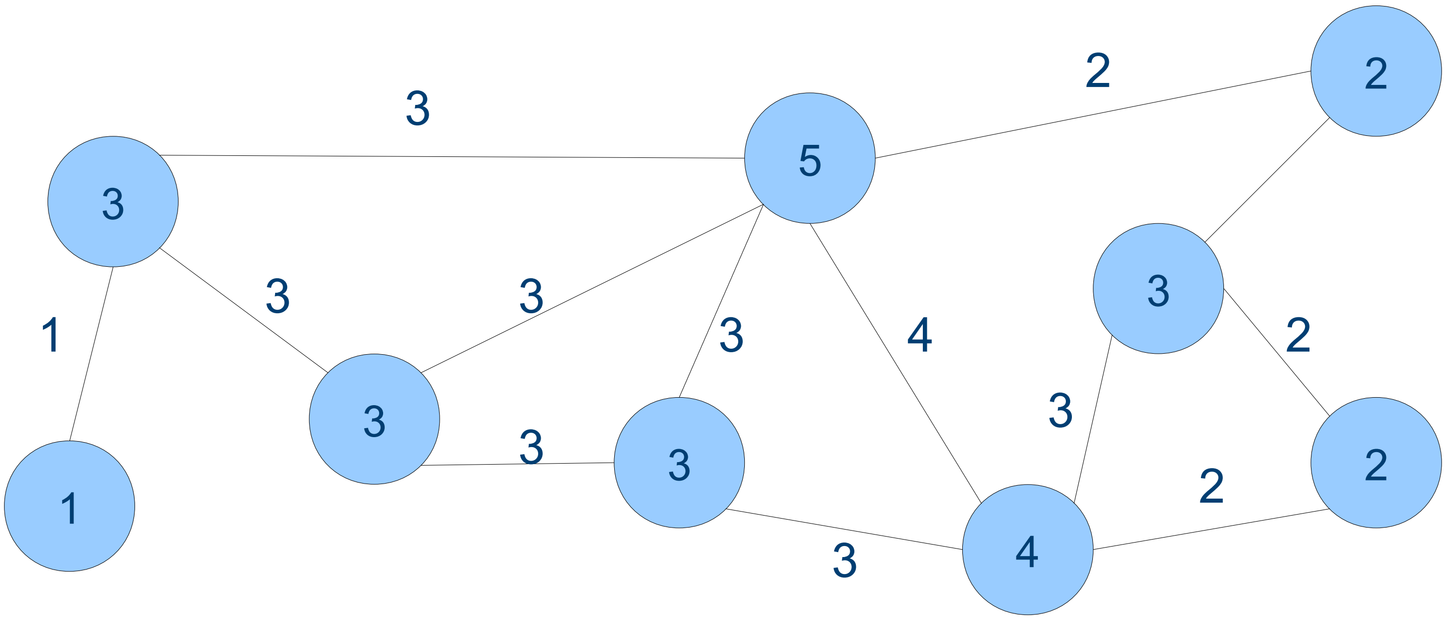
Can we find a 2-regular subgraph?

Regular subgraph extraction



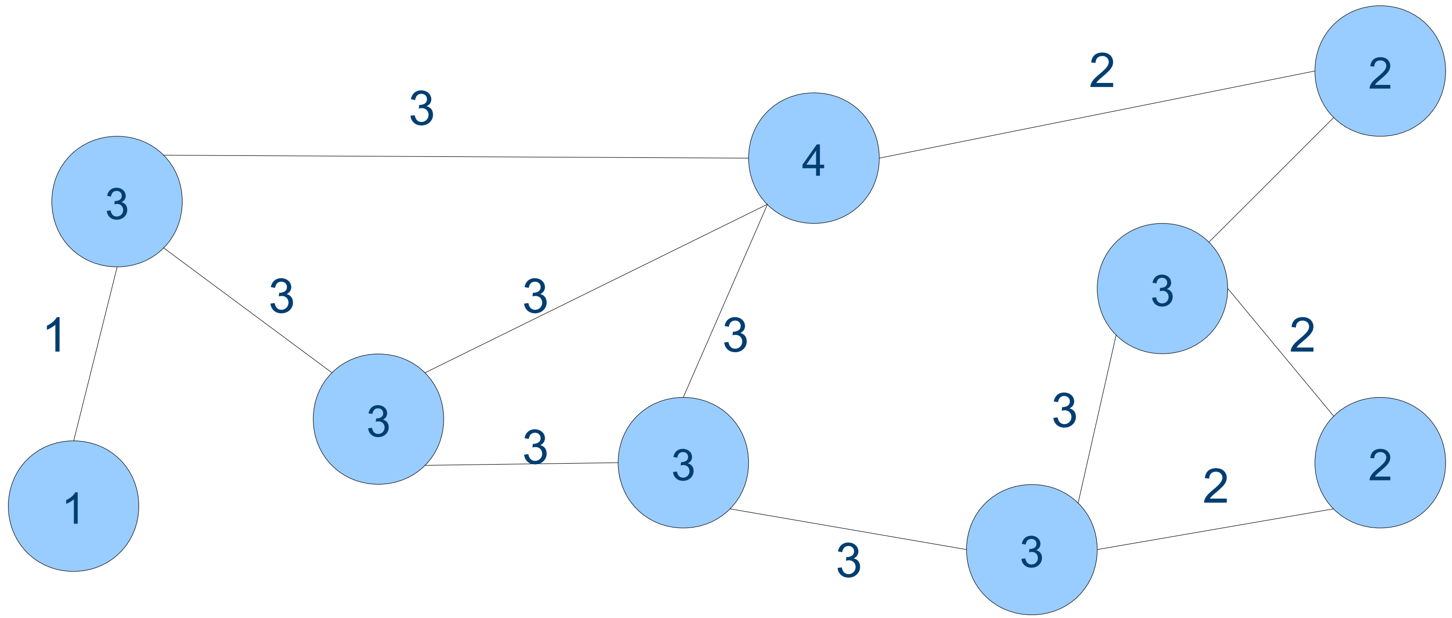
Step 1: Remove edges, weight by smallest attached node

Regular subgraph extraction



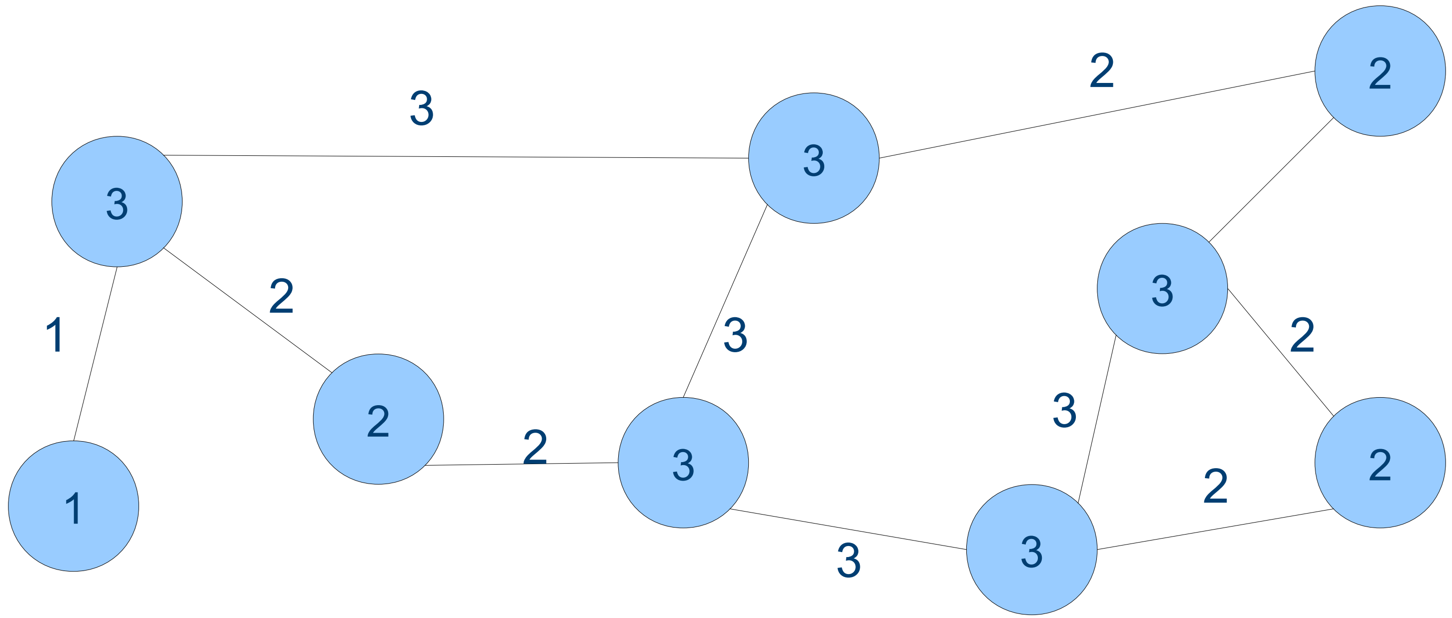
Step 1: Remove edges, weight by smallest attached node

Regular subgraph extraction



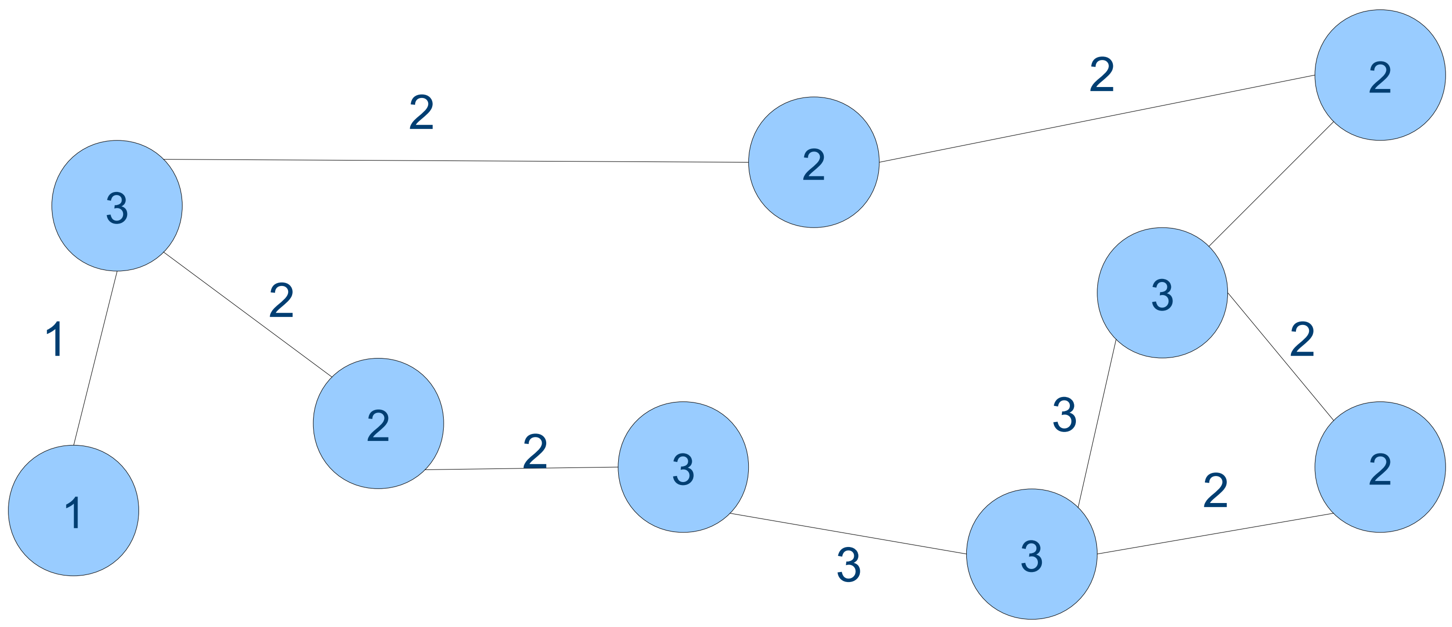
Step 1: Remove edges, weight by smallest attached node

Regular subgraph extraction



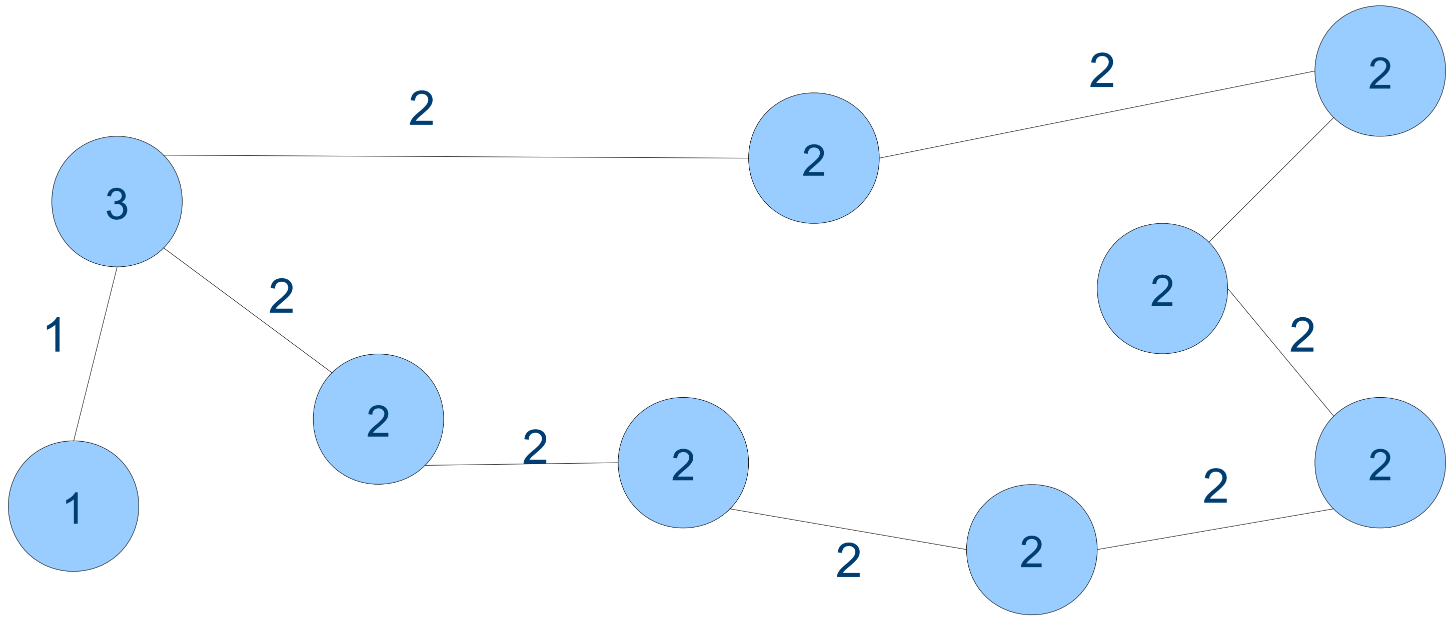
Step 1: Remove edges, weight by smallest attached node

Regular subgraph extraction



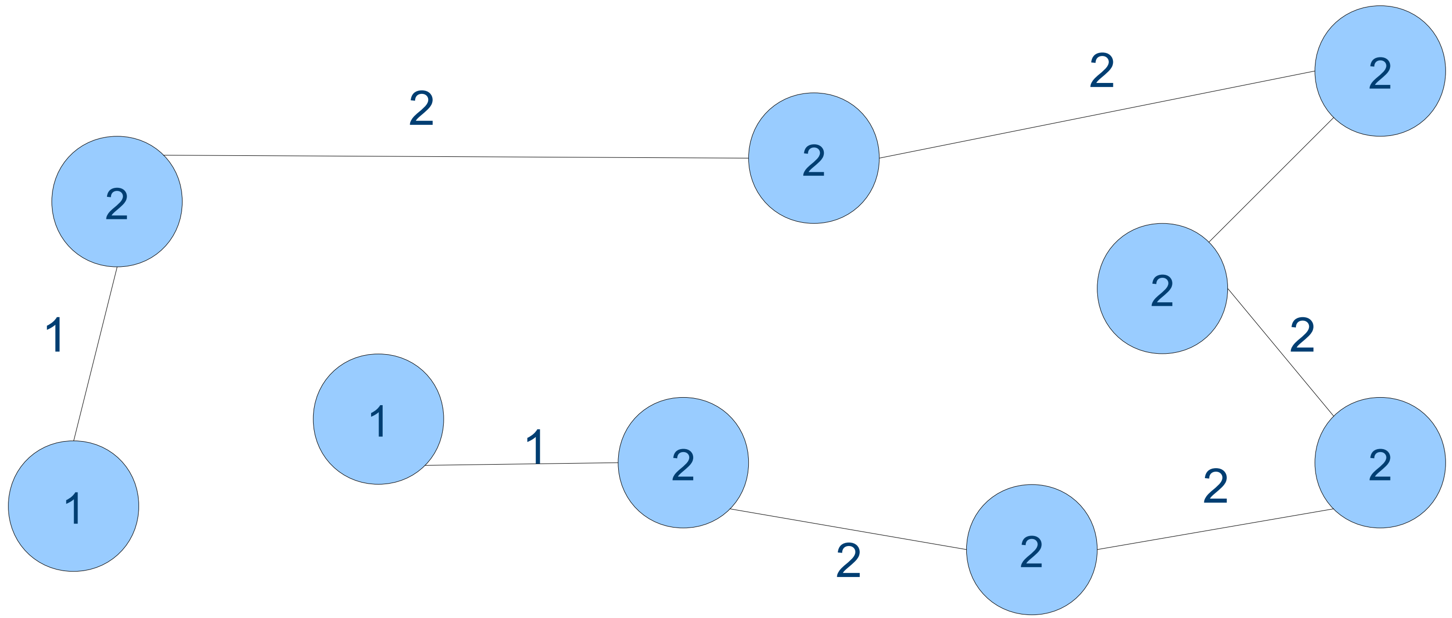
Step 1: Remove edges, weight by smallest attached node

Regular subgraph extraction



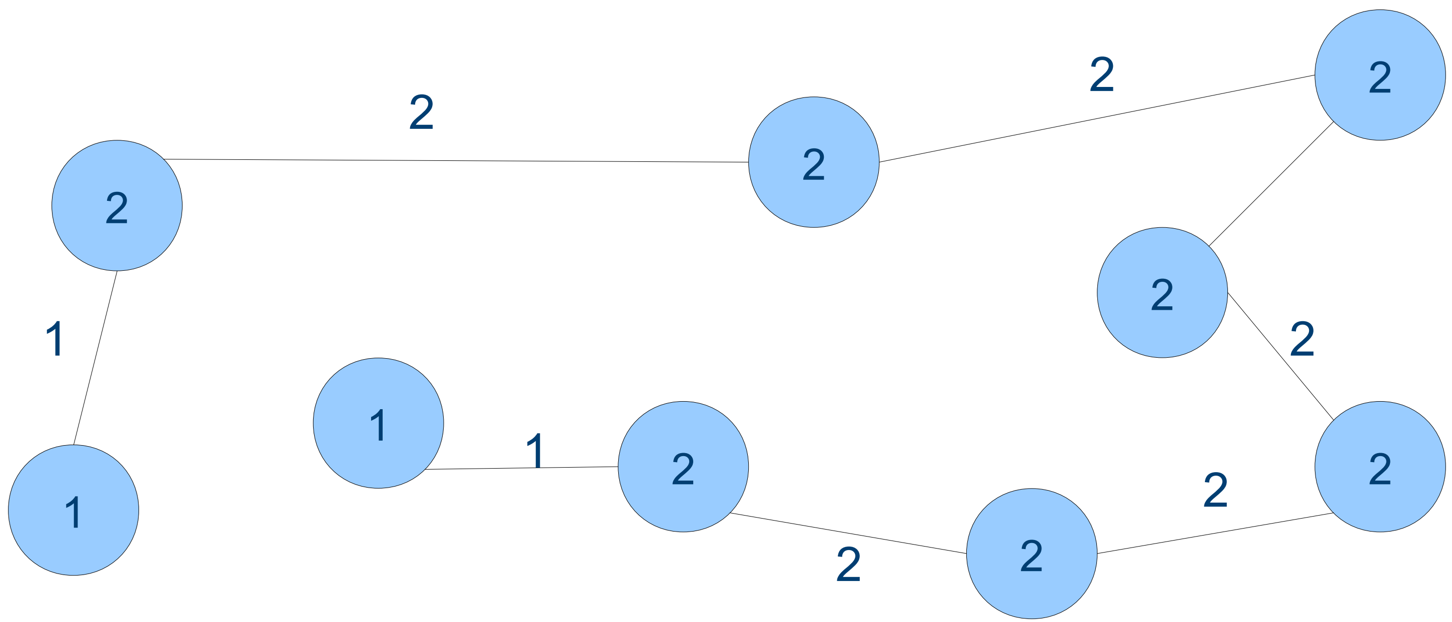
Step 2: Remove further edges to force all degrees ≤ k

Regular subgraph extraction



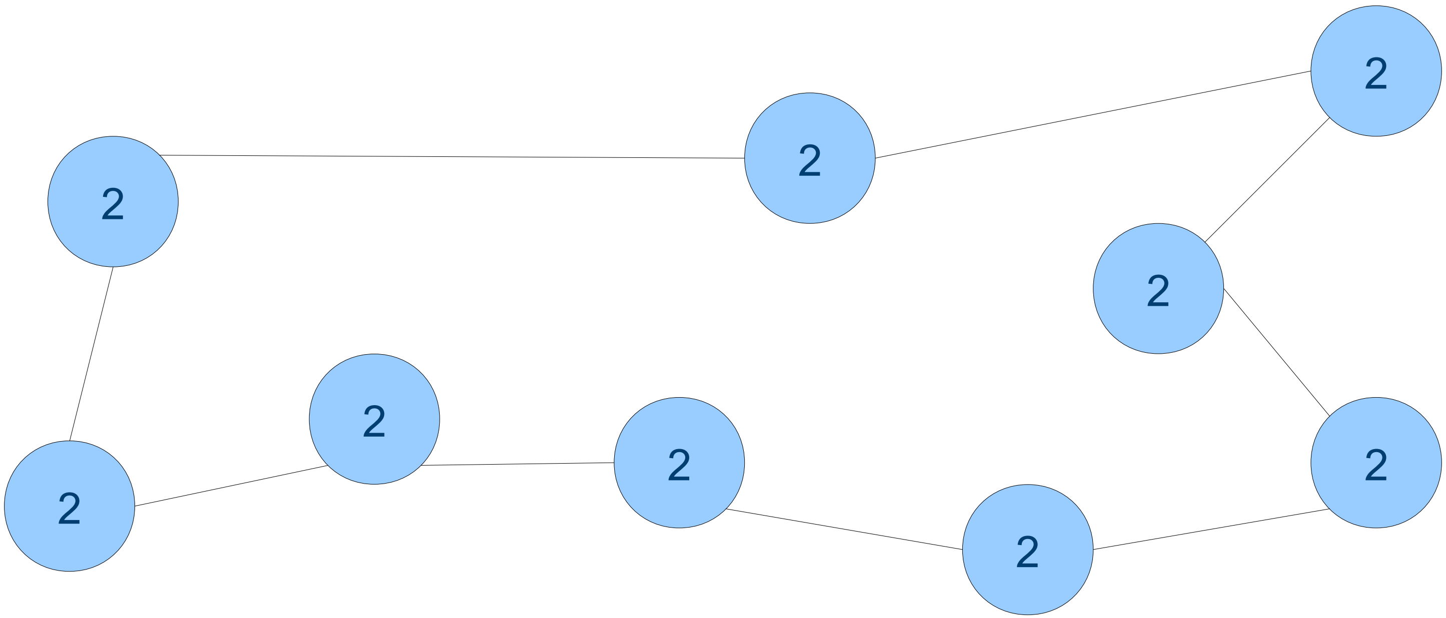
Step 3: Randomly add edges between pairs of edges below *k*

Regular subgraph extraction



Step 3: Randomly add edges between pairs of edges below *k*

Regular subgraph extraction



(note: producing a cycle is atypical!)

How well have we done?

- Recall original goal of showing *k*-sample
  - Promotion, identification

- Two measures:
  - *Precision:* Percentage of edges shown which are real
  - *Recall*: Percentage of real edges which are shown

    (normalise recall to showing a max of *k* per node)

How well have we done?

- Recall original goal of showing *k*-sample
  - Promotion, identification

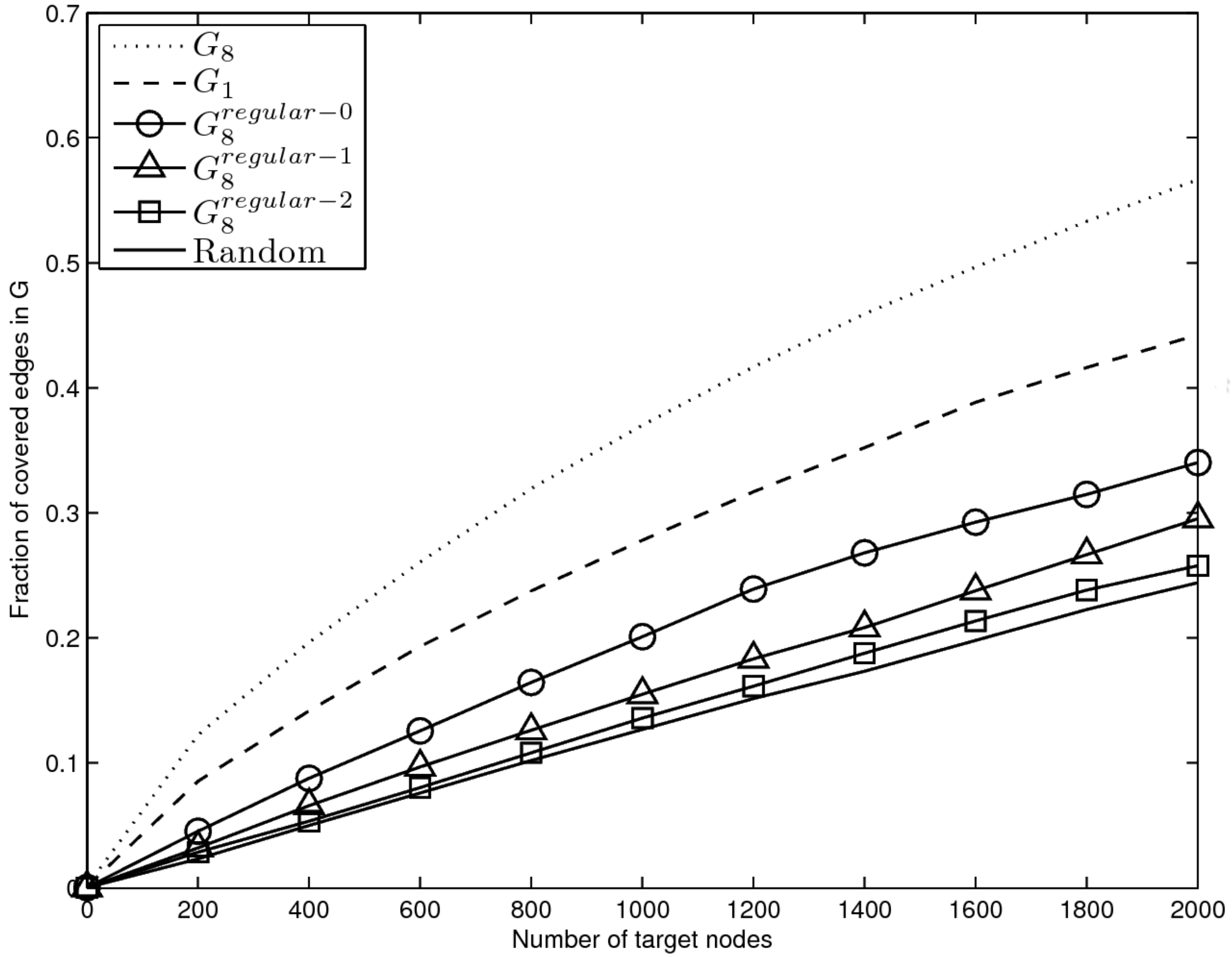
- Two measures:
  - *Precision:* Percentage of edges shown which are real
  - *Recall*: Percentage of real edges which are shown

    (normalise recall to showing a max of *k* per node)

Regular subgraph extraction

	Original	Step 1	Step 2	Step 3
Precision	1	1	1	0.90
Recall	1	1	0.99	0.99

Regular subgraph extraction



Drawbacks

- Requires complete graph knowledge
- Graph frequently changes!

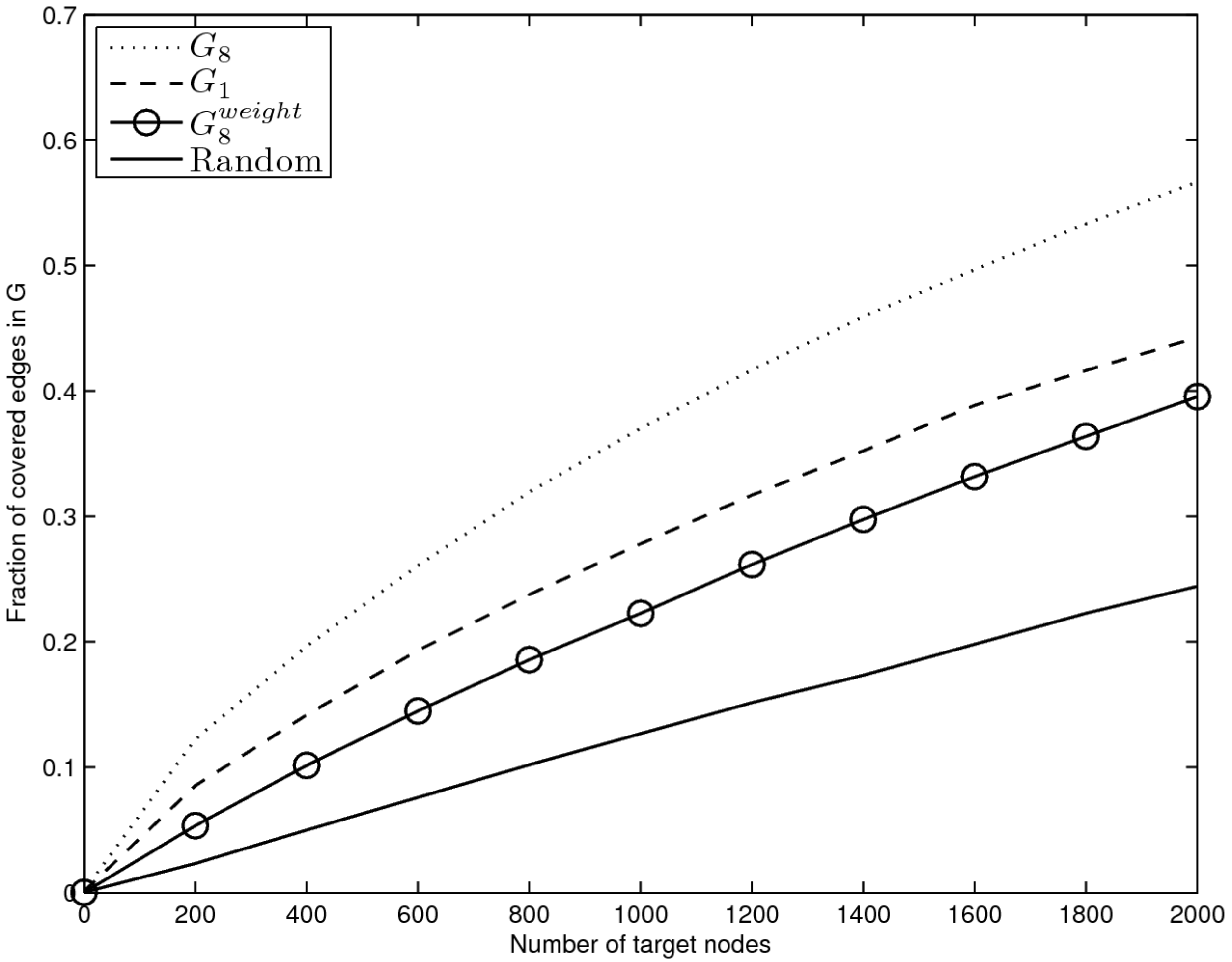
Drawbacks

- Requires complete graph knowledge
- Graph frequently changes!

Alternative: Random Sampling

- Weight selection towards low-degree neighbours
- Computable locally, incrementally
- (much weaker...)

Random Sampling



Caveats

- Can gain some protection against degree estimation
  - With a lot of work
- Doesn't prevent inference of dominating sets, centrality!

Conclusions

- Availability of social graphs raises serious privacy concern
  - The blueprint of our society...
- Very fragile to many attacks
- Right now, we're choosing utility over privacy

Thank You!

jcb82@cl.cam.ac.uk