UNIVERSITY OF TWENTE.

Finding central nodes in large networks

Nelly Litvak

University of Twente
Eindhoven University of Technology,
The Netherlands

Woudschoten Conference 2017

Complex networks

- Networks: Internet, WWW, social networks, neural networks,...

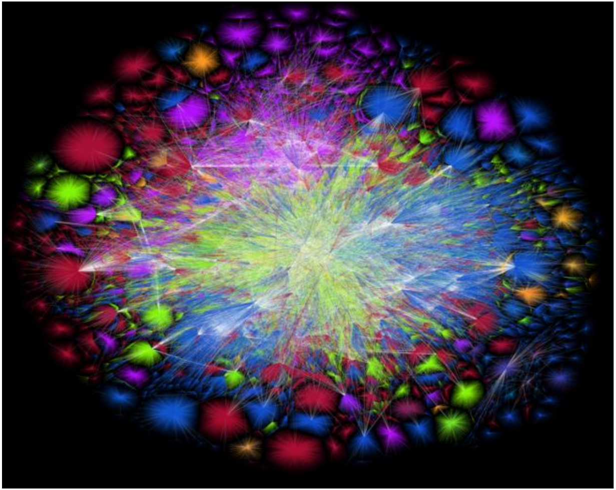
Complex networks

- Networks: Internet, WWW, social networks, neural networks,...
- Many nodes connected by edges

Complex networks

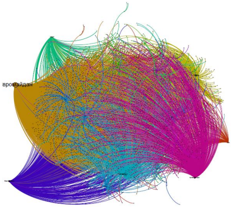
- Networks: Internet, WWW, social networks, neural networks,...
- Many nodes connected by edges
- Physics, computer science, sociology, biology, art

The Internet



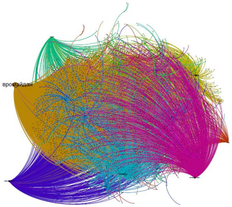
www.opte.org

Examples of networks

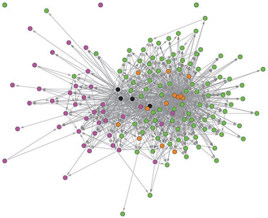


Euromaidan Retweets

Examples of networks



Euromaidan Retweets



Bank transactions
in Australia.

Centrality in networks

- Network as a graph $G = (V, E)$

Centrality in networks

- Network as a graph $G = (V, E)$
- Undirected:

Centrality in networks

- Network as a graph $G = (V, E)$
- Undirected: Facebook, roads networks, airline networks, power grids, co-authorship networks

Centrality in networks

- Network as a graph $G = (V, E)$
- Undirected: Facebook, roads networks, airline networks, power grids, co-authorship networks
- Directed:

Centrality in networks

- Network as a graph $G = (V, E)$
- Undirected: Facebook, roads networks, airline networks, power grids, co-authorship networks
- Directed: Twitter, bank transactions, food webs, WWW, scientific citations

Centrality in networks

- Network as a graph $G = (V, E)$
- Undirected: Facebook, roads networks, airline networks, power grids, co-authorship networks
- Directed: Twitter, bank transactions, food webs, WWW, scientific citations
- V set of vertices, E set of edges

Centrality in networks

- Network as a graph $G = (V, E)$
- Undirected: Facebook, roads networks, airline networks, power grids, co-authorship networks
- Directed: Twitter, bank transactions, food webs, WWW, scientific citations
- V set of vertices, E set of edges
- $|V| = n$, we can let $n \to \infty$

Centrality in networks

- Network as a graph $G = (V, E)$
- Undirected: Facebook, roads networks, airline networks, power grids, co-authorship networks
- Directed: Twitter, bank transactions, food webs, WWW, scientific citations
- V set of vertices, E set of edges
- $|V| = n$, we can let $n \to \infty$

- Which nodes are most ‘central’ in a network?

Google search

Brin and Page 1998 *The anatomy of a large-scale hypertextual web search engine.*

Page, Brin, Motwani and Winograd 1999 *The PageRank Citation Ranking: Bringing Order to the Web.*

Google search

Brin and Page 1998 *The anatomy of a large-scale hypertextual web search engine.*

Page, Brin, Motwani and Winograd 1999 *The PageRank Citation Ranking: Bringing Order to the Web.*

- Yahoo, AltaVista,...

Google search

Brin and Page 1998 *The anatomy of a large-scale hypertextual web search engine.*

Page, Brin, Motwani and Winograd 1999 *The PageRank Citation Ranking: Bringing Order to the Web.*

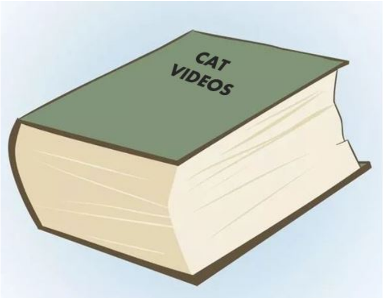
- Yahoo, AltaVista,...
- Directory-based, comparable to telephone books

Google search

Brin and Page 1998 *The anatomy of a large-scale hypertextual web search engine.*

Page, Brin, Motwani and Winograd 1999 *The PageRank Citation Ranking: Bringing Order to the Web.*

- Yahoo, AltaVista,...
- Directory-based, comparable to telephone books



Google PageRank

- PageRank r_i of page $i = 1, \ldots, n$ is defined as:

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1 - \alpha) q_i, \quad i = 1, \ldots, n$$

- $d_j = \#$ out-links of page j
- $\alpha \in (0, 1)$, *damping factor* originally 0.85
- $q_i \geqslant 0$, $\sum_i q_i = 1$, originally, $q_i = 1/n$.

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1 - \alpha$ random jump (to page i w.p. q_i)

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1-\alpha$ random jump (to page i w.p. q_i)
- Dangling nodes, $d_j = 0$:
  - Random jump from dangling nodes

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1-\alpha$ random jump (to page i w.p. q_i)
- Dangling nodes, $d_j = 0$:
  - Random jump from dangling nodes
  - Stationary distribution $\pi = \mathbf{r}/\|\mathbf{r}\|_1$

Easily bored surfer model

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

- r_i is a stationary distribution of a Markov chain
- with probability α follow a randomly chosen outgoing link
- with probability $1-\alpha$ random jump (to page i w.p. q_i)
- Dangling nodes, $d_j = 0$:
  - Random jump from dangling nodes
  - Stationary distribution $\pi = \mathbf{r}/\|\mathbf{r}\|_1$
- The page is important if many important pages link to it!

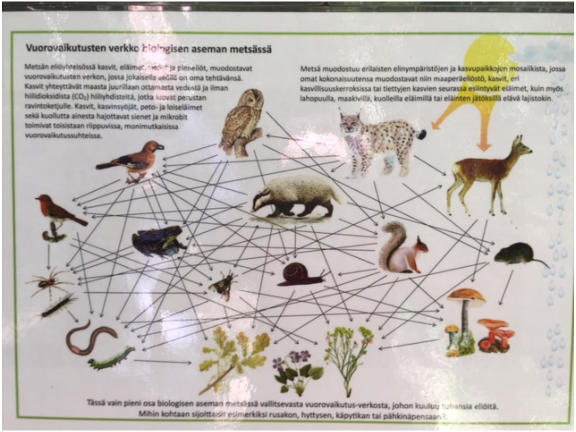
PageRank beyond web search

- Applications:
  - Topic-sensitive search (Haveliwala 2002);
  - Spam detection (Gyöngyi et al. 2004)
  - Finding related entities (Chakrabarti 2007);
  - Link prediction (Liben-Nowell and Kleinberg 2003; Voevodski, Teng, Xia 2009);
  - Finding local cuts (Andersen, Chung, Lang 2006);
  - Graph clustering (Tsiatas, Chung 2010);
  - Person name disambiguation (Smirnova, Avrachenkov, Trousse 2010);
  - Finding most influential people in Wikipedia (Shepelyansky et al 2010, 2013)

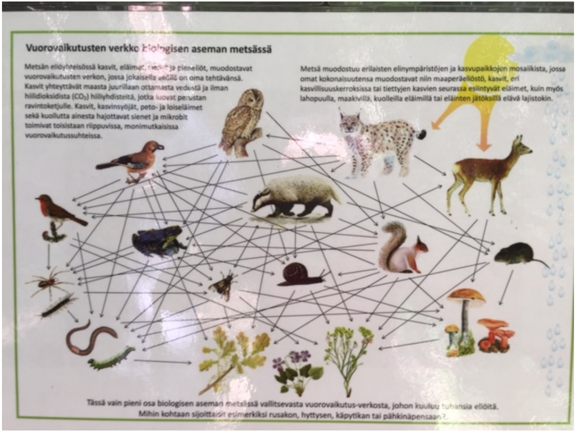
PageRank beyond web search

- Applications:
  - Topic-sensitive search (Haveliwala 2002);
  - Spam detection (Gyöngyi et al. 2004)
  - Finding related entities (Chakrabarti 2007);
  - Link prediction (Liben-Nowell and Kleinberg 2003; Voevodski, Teng, Xia 2009);
  - Finding local cuts (Andersen, Chung, Lang 2006);
  - Graph clustering (Tsiatas, Chung 2010);
  - Person name disambiguation (Smirnova, Avrachenkov, Trousse 2010);
  - Finding most influential people in Wikipedia (Shepelyansky et al 2010, 2013)
- Global characteristic of the graph

Example: food web



Example: food web



Allesina and Pascual 2009

Matrix form

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

$$P = \begin{cases} \frac{1}{d_j}, & j \to i \\ 0, & \textit{otherwise}. \end{cases}$$

$\mathbf{r} = (r_1, r_2, \ldots, r_n)$
$\mathbf{q} = (q_1, q_2, \ldots, q_n)$

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

Matrix form

$$r_i = \sum_{j\,:\,j \to i} \frac{\alpha}{d_j} r_j + (1-\alpha) q_i, \quad i = 1, \ldots, n$$

$$P = \begin{cases} \frac{1}{d_j}, & j \to i \\ 0, & \textit{otherwise}. \end{cases}$$

$\mathbf{r} = (r_1, r_2, \ldots, r_n)$
$\mathbf{q} = (q_1, q_2, \ldots, q_n)$

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

Linear equation and eigenvector problem

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = \mathbf{r}\left[\alpha P + \frac{1-\alpha}{n}\mathbf{1}^t\mathbf{q}\right] \qquad \text{eigenvector problem}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I-\alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty}\alpha^t P^t.$$

Matrix expansion

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty} \alpha^t P^t.$$

- Computation by matrix iterations:

$$\begin{aligned}
\mathbf{r}^{(0)} &= (1/n, \ldots, 1/n) \\
\mathbf{r}^{(k)} &= \alpha \mathbf{r}^{(k-1)} P + (1-\alpha)\mathbf{q} \\
&= \mathbf{r}^{(0)} \alpha^k P^k + (1-\alpha)\mathbf{q}\sum_{t=0}^{k-1} \alpha^t P^t
\end{aligned}$$

Matrix expansion

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q} \sum_{t=0}^{\infty} \alpha^t P^t.$$

- Computation by matrix iterations:

$$\begin{aligned}
\mathbf{r}^{(0)} &= (1/n, \ldots, 1/n) \\
\mathbf{r}^{(k)} &= \alpha \mathbf{r}^{(k-1)} P + (1-\alpha)\mathbf{q} \\
&= \mathbf{r}^{(0)} \alpha^k P^k + (1-\alpha)\mathbf{q} \sum_{t=0}^{k-1} \alpha^t P^t
\end{aligned}$$

- Exponentially fast convergence due to $\alpha \in (0, 1)$

Matrix expansion

$$\mathbf{r} = \alpha \mathbf{r} P + (1-\alpha)\mathbf{q}$$

$$\mathbf{r} = (1-\alpha)\mathbf{q}[I - \alpha P]^{-1} = (1-\alpha)\mathbf{q}\sum_{t=0}^{\infty} \alpha^t P^t.$$

- Computation by matrix iterations:

$$\begin{aligned} \mathbf{r}^{(0)} &= (1/n, \ldots, 1/n) \\ \mathbf{r}^{(k)} &= \alpha \mathbf{r}^{(k-1)} P + (1-\alpha)\mathbf{q} \\ &= \mathbf{r}^{(0)} \alpha^k P^k + (1-\alpha)\mathbf{q} \sum_{t=0}^{k-1} \alpha^t P^t \end{aligned}$$

- Exponentially fast convergence due to $\alpha \in (0, 1)$
- Matrix iterations are used to compute PageRank in practice Langville&Meyer 2004, Berkhin 2005

Ranking algorithms/Centrality measures

Recent review: (Boldi and Vigna 2014)

- Based on distances:
  - (in-)degree: number of nodes on distance 1
  - Closeness centrality (Bavelas 1950)
  - Harmonic centrality (Boldi and Vigna 2014)
- Based on paths:
  - Betweenness centrality (Anthonisse 1971)
  - Katz's index (Katz 1953)
- Based ob spectrum:
  - Seeley index (Seeley 1949)
  - HITS (Kleinberg 1997)
  - PageRank (Brin, Page, Motwani and Vinograd 1999)

Plan

- Part I: Centrality & computaitonal aspects
- Part II: PageRank

Degree

- The degree of a node is a number of edges attached to it

Degree

- The degree of a node is a number of edges attached to it
- Directed graph: in- and out-degree

Degree

- The degree of a node is a number of edges attached to it
- Directed graph: in- and out-degree
- Is (in-)degree a good centrality measure?

Degree

- The degree of a node is a number of edges attached to it
- Directed graph: in- and out-degree
- Is (in-)degree a good centrality measure?
- Easy to compute?

Finding top-k most followed users on Twitter

- Problem: Find top-k network nodes with most number of connections

Finding top-k most followed users on Twitter

- Problem: Find top-k network nodes with most number of connections
- Some applications:
  - Routing via large degree nodes
  - Proxy for various centrality measures
  - Node clustering and classification
  - Epidemic processes on networks
  - Finding most popular entities (e.g. interest groups)

Finding top-k most followed users on Twitter

- Problem: Find top-k network nodes with most number of connections
- Some applications:
  - Routing via large degree nodes
  - Proxy for various centrality measures
  - Node clustering and classification
  - Epidemic processes on networks
  - Finding most popular entities (e.g. interest groups)
  - Many companies maintain network statistics (*twittercounter.com*, *followerwonk.com*, *twitaholic.com*, *www.insidefacebook.com*, *yavkontakte.ru*)

Finding top-k largest degree nodes

- If the information about a complete network is available, complexity $O(n)$

Finding top-k largest degree nodes

- If the information about a complete network is available, complexity $O(n)$
- Twitter has one billion accounts

Finding top-k largest degree nodes

- If the information about a complete network is available, complexity $O(n)$
- Twitter has one billion accounts
- The network can be accessed only via API, one access per minute. It will take 900 years to crawl Twitter!

Finding top-k largest degree nodes

- If the information about a complete network is available, complexity $O(n)$
- Twitter has one billion accounts
- The network can be accessed only via API, one access per minute. It will take 900 years to crawl Twitter!
- Randomized algorithms: Find a 'good enough' answer with a small number of API requests.

Known algorithms

- Random-walk based. Cooper, Radzik, Siantos 2012
  Transitions probabilities along undirected edges (i, j) are proportional to $(d(i)d(j))^b$, where $d(i)$ is the degree of a vertex x and $b > 0$ is some parameter.
- Random Walk Avrachenkov, L, Sokol, Towsley 2012 Random walk with uniform jumps. In an undirected graphs the stationary distribution is a linear function of degrees.
- Crawl-AI and Crawl-GAI. Kumar, Lang, Marlow, Tomkins 2008 At every step all nodes have their *apparent in-degrees* S_j, $j = 1, \ldots, n$: the number of discovered edges pointing to this node. Designed for WWW crawl.
- HighestDegree. Borgs, Brautbar, Chayes, Khanna, Lucier 2012 Retrieve a random node, check in-degrees of its out-neighbors. Proceed while resources are available
- Two-stage algorithm. Avrachenkov,L,Ostroumova 2014

The Friendship Paradox

- Feld, 1991

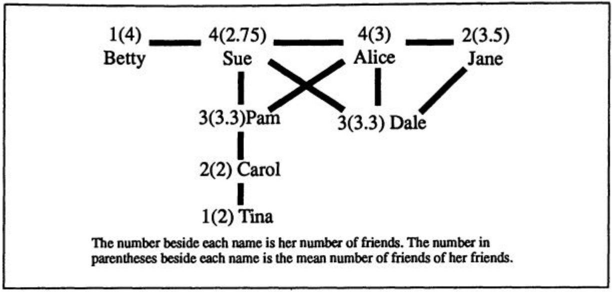


FIG. 1.—Friendships among eight girls at Marketville High School

The Friendship Paradox

- Feld, 1991

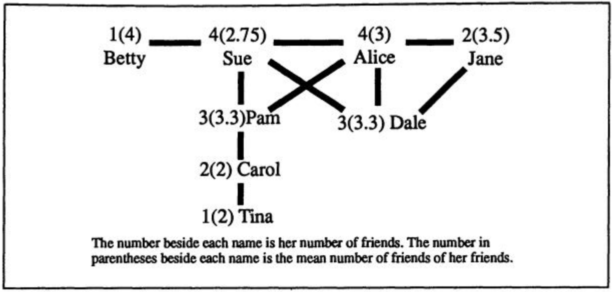


FIG. 1.—Friendships among eight girls at Marketville High School

- In the figure: # friends (average number of friends' friends)

The Friendship Paradox

- Feld, 1991

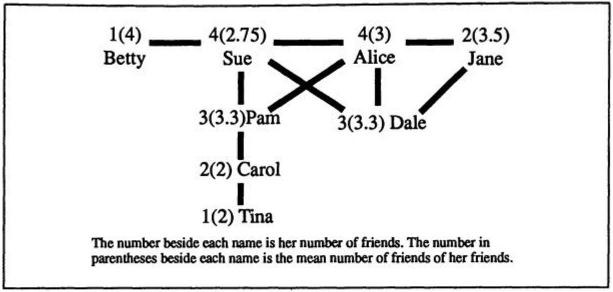


FIG. 1.—Friendships among eight girls at Marketville High School

- In the figure: # friends (average number of friends' friends)
- More popular than her friends: Sue, Alice
- As popular as her friends: Carol
- Less popular than her friends: Betty, Pam, Tina, Dale, Jane

Friendship paradox

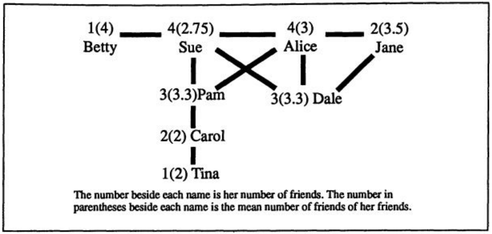


FIG. 1.—Friendships among eight girls at Marketville High School

- People with may connections are more likely to be your friend

Friendship paradox

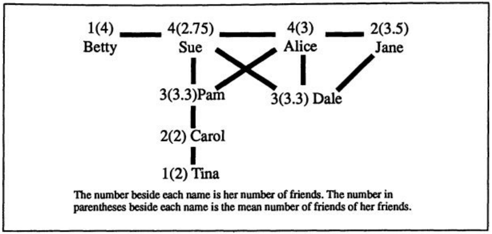


FIG. 1.—Friendships among eight girls at Marketville High School

- People with may connections are more likely to be your friend
- Sampling by edge is biased towards nodes with high degrees

Friendship paradox

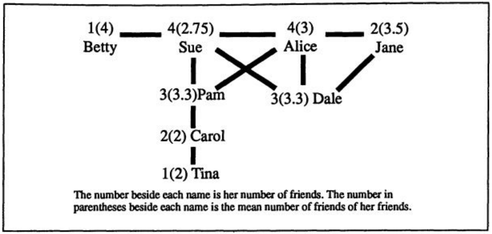
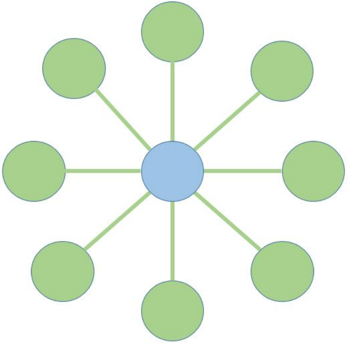


FIG. 1.—Friendships among eight girls at Marketville High School

- People with may connections are more likely to be your friend
- Sampling by edge is biased towards nodes with high degrees

Friendship paradox: star graph



Friendship paradox: consequences

- Friendship paradox can lead to wrong sampling

Friendship paradox: consequences

- Friendship paradox can lead to wrong sampling
- But it can be also exploited!

Friendship paradox: consequences

- Friendship paradox can lead to wrong sampling
- But it can be also exploited!
- Neighbor vaccinations

Friendship paradox: consequences

- Friendship paradox can lead to wrong sampling
- But it can be also exploited!
- Neighbor vaccinations
- Most followed Twitter users

Friendship paradox: consequences

- Friendship paradox can lead to wrong sampling
- But it can be also exploited!
- Neighbor vaccinations
- Most followed Twitter users

	Twitter users	Followers	Following	Tweets
1	KATY PERRY @katyperry	95,607,996	190	7,608
2	Justin Bieber @justinbieber	91,539,626	300,977	30,645
3	Barack Obama @BarackObama	84,088,937	631,665	15,434
4	Taylor Swift @taylorswift13	83,302,469	244	4,161
5	Rihanna @rihanna	69,480,199	1,134	9,898
6	YouTube @YouTube	66,399,134	986	18,768

Exploiting the Friendship Paradox

Exploiting the Friendship Paradox

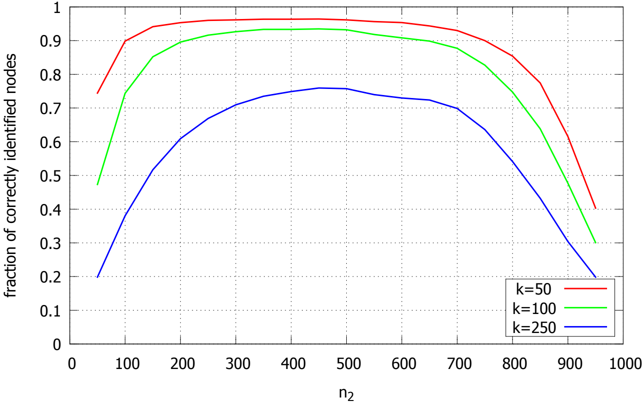
- Step 1: Select N_1 random users, see whom they follow (N_1 API requests)
- Friendship paradox: the people that a random user follows are often the most popular users in the network

Exploiting the Friendship Paradox

- Step 1: Select N_1 random users, see whom they follow (N_1 API requests)
- Friendship paradox: the people that a random user follows are often the most popular users in the network
- Step 2: Check, say, N_2 accounts, most followed by the group of N_1 random users chosen in Step 1. Top-k accounts should be there with high probability!

In total, we use $N_1 + N_2 = N$ requests to API

Results on Twitter



The fraction of correctly identified top-k most followed Twitter users. Horisonal axis: number of requests in Step 2. Total number of requests is $N = 1000$.

Comparison to known algorithms

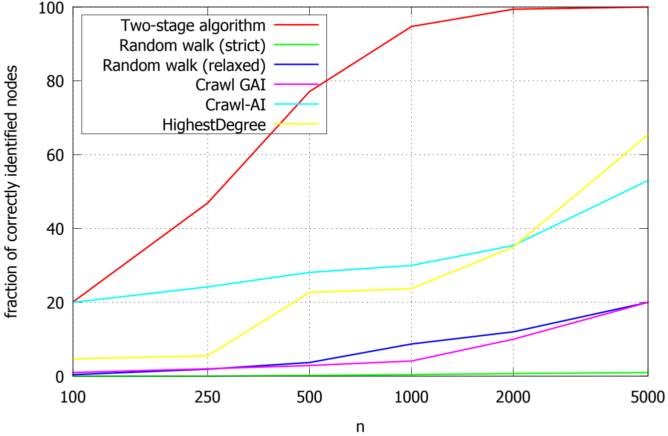


Figure: The fraction of correctly identified top-100 most followed Twitter users as a function of the number of API averaged over 10 experiments.

Advantages of the two-stage algorithm

- Does not waste resources
- Obtains *exact* degrees of the N_2 'most promising' nodes

Hubs in complex networks

- D is (in-)degree of a random node
- Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \tag{1}$$

$L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$

Hubs in complex networks

- D is (in-)degree of a random node
- Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \tag{1}$$

  $L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$
- 'Scale-free' distribution
- Some nodes (hubs) have really high degrees

Hubs in complex networks

- D is (in-)degree of a random node
- Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \tag{1}$$

$L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$

- 'Scale-free' distribution
- Some nodes (hubs) have really high degrees
- Top-degrees are top order statistics

Hubs in complex networks

- D is (in-)degree of a random node
- Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \tag{1}$$

  $L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$
- 'Scale-free' distribution
- Some nodes (hubs) have really high degrees
- Top-degrees are top order statistics
- Extreme value theory
  - Top-k order degrees 'of the order' $n^{1/\gamma}k^{1/\gamma}$

Hubs in complex networks

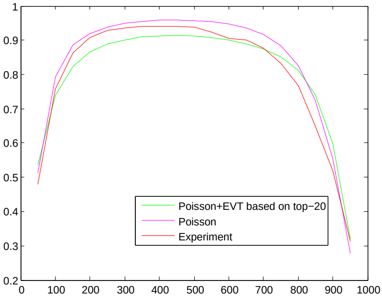
- D is (in-)degree of a random node
- Regular varying distribution:

$$P(D > x) = L(x)x^{-\gamma} \quad (1)$$

  $L(x)$ is slowly varying, i.e. $\lim_{t\to\infty} L(tx)/L(t) = 1$, $x \geqslant 0$
- 'Scale-free' distribution
- Some nodes (hubs) have really high degrees
- Top-degrees are top order statistics
- Extreme value theory
  - Top-k order degrees 'of the order' $n^{1/\gamma}k^{1/\gamma}$
  - Heurostic 'proof': $P(D > x) \approx k/n$

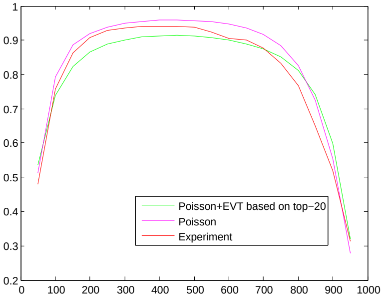
Performance evaluation

- Sublinear complexity $N = O(n^{1-1/\gamma})$
- Prediction of the performance of the algorithm



Performance evaluation

- Sublinear complexity $N = O(n^{1-1/\gamma})$
- Prediction of the performance of the algorithm



- Network sampling
  - vaccinations (L&Holme 2017), marketing, P2P

Katz's index

Katz (1953)

$\mathbf{1}$ – vector of ones

- Classical version of PageRank

$$n\mathbf{r} = (1-\alpha)\mathbf{1}[I-\alpha P]^{-1},$$

  P is a matrix of a simple random walk on the graph

- Katz's index

$$\mathbf{k} = (1-\beta)\mathbf{1}[I-\beta A]^{-1},$$

  A - adjacency matrix of the graph

- $\beta < 1/\lambda$, where λ is the dominant eigenvalue of A

Closeness centrality

- $d(i, j)$ – graph distance between i and j
- no path, then $d(i, j) = \infty$
- Closeness centrality of node i

$$\frac{1}{\sum_{j:d(i,j)<\infty} d(i, j)}$$

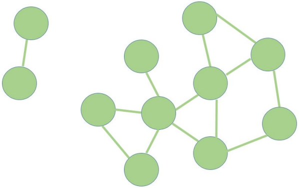
- Problem?

Closeness centrality

- $d(i, j)$ – graph distance between i and j
- no path, then $d(i, j) = \infty$
- Closeness centrality of node i

$$\frac{1}{\sum_{j:d(i,j)<\infty} d(i, j)}$$

- Problem?



- Maximal closeness for two disconnected vertices
- How do we compute distances?

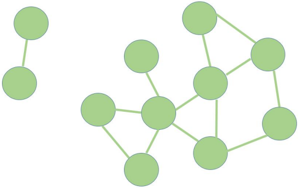
Harmonic centrality

Boldi& Vigna (2014)

- Closeness centrality $\frac{1}{\sum_{j:d(i,j)<\infty} d(i,j)}$
- Harmonic centrality

$$\sum_{j \neq i} \frac{1}{d(i,j)}$$

- Maximal for central nodes in large components



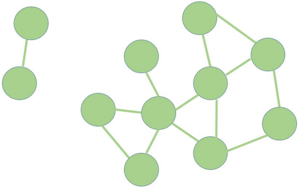
Harmonic centrality

Boldi& Vigna (2014)

- Closeness centrality $\frac{1}{\sum_{j:d(i,j)<\infty} d(i,j)}$
- Harmonic centrality

$$\sum_{j \neq i} \frac{1}{d(i,j)}$$

- Maximal for central nodes in large components



- HyperLogLog-type algorithm to compute distances

Betweenness centrality

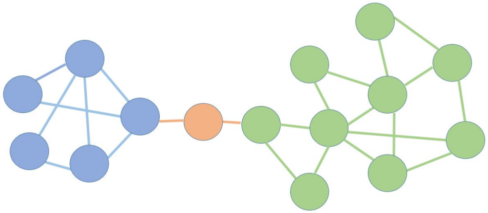
- σ_{st} – number of shortest paths from s to t
- $\sigma_{st}(i)$ – number of shortest paths from s to t through i

Betweenness centrality

- σ_{st} – number of shortest paths from s to t
- $\sigma_{st}(i)$ – number of shortest paths from s to t through i
- Betweenness centrality of i

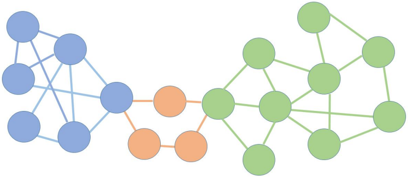
$$\sum_{s,t \neq i, \sigma_{s,t} \neq 0} \frac{\sigma_{st}(i)}{\sigma_{st}}.$$

- Fraction of shortest paths through i



Current flow betweenness centrality

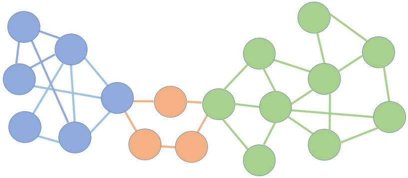
Newman (2005)



- $V_i(s,t)$ – # visits to i of a random walk from s to t
- $|V_j(s,t) - V_i(s,t)|$ – centrality of edge $\{i,j\}$
- Current through $\{i,j\}$ when 1 unit current goes from s to t
- Complexity $O((I(n1) + n \log n |E|)$

Current flow betweenness centrality

Newman (2005)



- $V_i(s,t)$ – # visits to i of a random walk from s to t
- $|V_j(s,t) - V_i(s,t)|$ – centrality of edge $\{i,j\}$
- Current through $\{i,j\}$ when 1 unit current goes from s to t
- Complexity $O((I(n1) + n\log n|E|)$
- Avrachenkov, L, Medyanikov, Sokol (2013) α-current flow betweenness centrality
- At each step the random walk continues with probability α
- Similar to PageRank

Open Wikipedia ranking

http://wikirank.di.unimi.it/