An Introduction to Social Mining

Vladimir Gorovoy* and Yana Volkovich†

†@yvolkovich
Barcelona Media, Information, Technology & Society Group
Barcelona, Spain

* @vgorovoy
Yandex, Yandex.Uslugi
Saint Petersburg, Russia

August, 15-19 2011

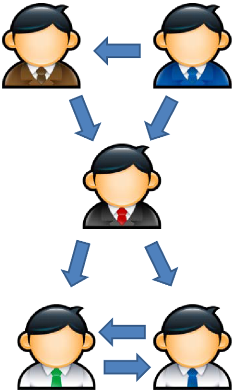
Outline

1. Graph mining
   - Graph construction
   - Matrix analysis
   - Power law
   - Small-world effect
   - Assortativity
2. Influence propagation

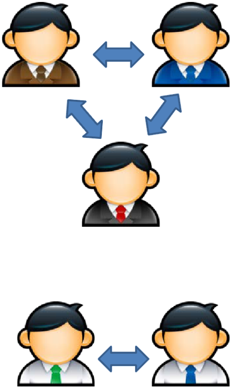
- Social Media $\rightarrow$ social presence, social interactions
- graph **G=G(V,E)**,
  - *V* is the set of vertices, or nodes,
  - *E* is the set of edges (edges may have weights)

Graph mining
Example

- **'user $\rightleftarrows$ user'** graphs on the base of social interactions (e.g. friendship, communications: sharing, commenting)



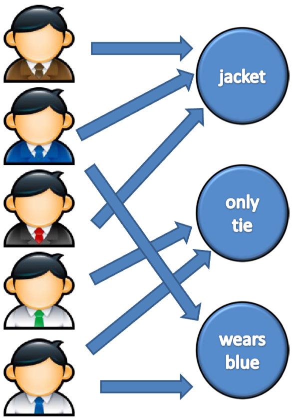
- **'user $\rightleftarrows$ user'** graphs on the base of social interactions (e.g. friendship, communications: sharing, commenting)



Graph mining

Example

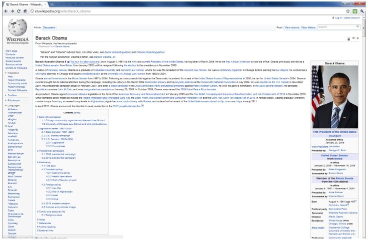
- **'user $\rightleftarrows$ properties'** bipartite graphs



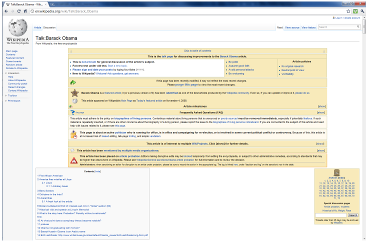
Graph mining

Example: Wikipedia

- Social interactions on Wikipedia [Laniado et al., 2011]
- *hidden side* of Wikipedia
  - article talk pages → explicit coordination and discussion
  - user talk pages → personal communications (sort of *public inbox*)
- Article Barack Obama:
  - discussion split into 72 pages
  - 22 000 comments in the article talk pages (17 500 edits done to the article)



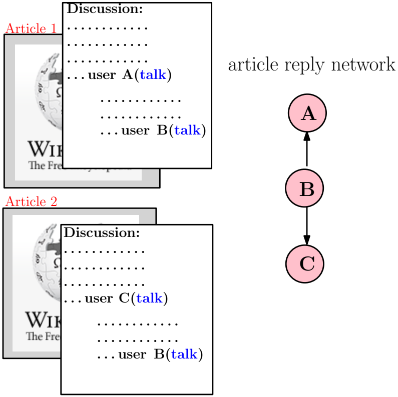
(c) article page



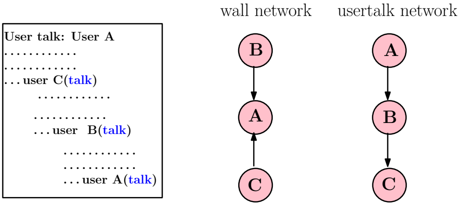
(d) discussion page

Example: Wikipedia graphs construction

- **article reply network** → direct replies in articles discussion pages.



- **user reply network** → direct replies in user talk pages.
- **wall network** → personal messages posted on another user's talk page.



- Jaccard coefficient of the overlap between the networks

$$C_{jaccard} = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} \cdot \frac{\max(|E_1|, |E_2|)}{\min(|E_1|, |E_2|)},$$

  - normalized to have a result in the interval [0,1]

	article-NW	**talk-NW**	**wall-NW**
article-NW	1	0.11	0.09
talk-NW	0.11	1	0.35
wall-NW	0.09	0.35	1

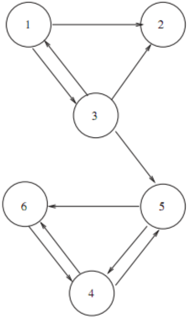
RuSSIR
Russian Summer School in Information Retrieval

Matrix analysis

- to represent a graph: Adjacency matrix

$$A = \{a_{i,j} |\ a_{i,j} = w_{i,j} \text{ iff } i \to j\};$$

- example from [Langville and Meyer, 2004]

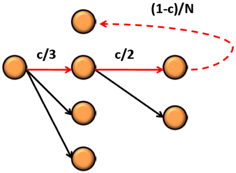


$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} \;\;1\;\; & \;\;2\;\; & \;\;3\;\; & \;\;4\;\; & \;\;5\;\; & \;\;6\;\; \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}. \end{array}$$

- **local characteristics**: in- and out-degrees, weighted degrees;
- **global characteristics**: PageRank and modifications;

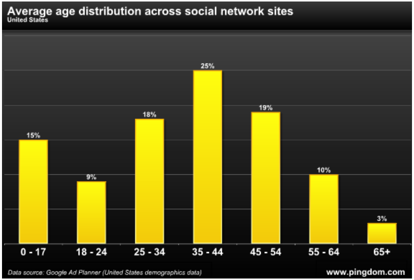
$$PR(i) = c \sum_{j \to i} \frac{1}{d_j} PR(j) + \frac{1-c}{N}.$$

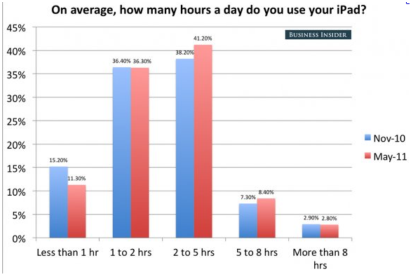
- stationary distribution of an '*easily-bored-surfer*' random walk on a graph

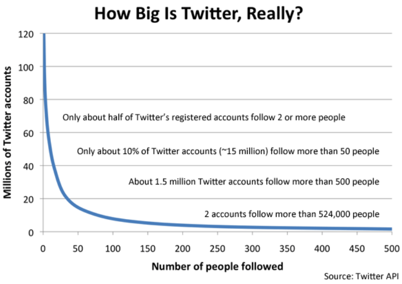


Power law

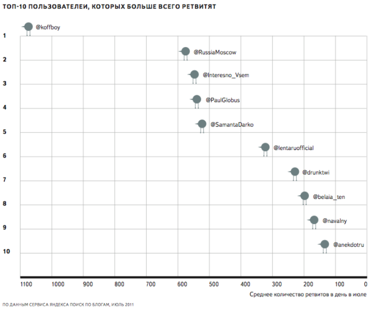
Quiz (1)


Average age distribution across social network sites
United States
15%
9%
18%
25%
19%
10%
3%
0 - 17
18 - 24
25 - 34
35 - 44
45 - 54
55 - 64
65+
Data source: Google Ad Planner (United States demographics data)
www.pingdom.com


On average, how many hours a day do you use your iPad?
BUSINESS INSIDER
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
15.20%
11.30%
36.40%
36.30%
38.20%
41.20%
7.30%
8.40%
2.90%
2.80%
Less than 1 hr
1 to 2 hrs
2 to 5 hrs
5 to 8 hrs
More than 8 hrs
Nov-10
May-11



How Big Is Twitter, Really?
Millions of Twitter accounts
120
100
80
60
40
20
0
Only about half of Twitter's registered accounts follow 2 or more people
Only about 10% of Twitter accounts (~15 million) follow more than 50 people
About 1.5 million Twitter accounts follow more than 500 people
2 accounts follow more than 524,000 people
0 50 100 150 200 250 300 350 400 450 500
Number of people followed
Source: Twitter API



ТОП-10 ПОЛЬЗОВАТЕЛЕЙ, КОТОРЫХ БОЛЬШЕ ВСЕГО РЕТВИТЯТ
1 @koffboy
2 @RussiaMoscow
3 @Interesno_Vsem
4 @PaulGlobus
5 @SamantaDarko
6 @lentaruofficial
7 @drunktwi
8 @belaia_ten
9 @navalny
10 @anekdotru
1100 1000 900 800 700 600 500 400 300 200 100 0
Среднее количество ретвитов в день в июле
ПО ДАННЫМ СЕРВИСА ЯНДЕКСА ПОИСК ПО БЛОГАМ, ИЮЛЬ 2011

What is the difference?

- Power law is a special family of distributions:
  - human heights, speed a car;
  - city population, # books sold, diameters of moon craters.

- random variable X has a **power law distribution** with exponent α:

$$\mathbb{P}(X > x) \sim x^{-\alpha} \text{ as } x \to \infty;$$

- **Pareto principle:** for many events roughly 80% of the effects come from 20% of the causes;
- α between 1 and 2: finite mean, infinite variance.

Power law

Log-log plot

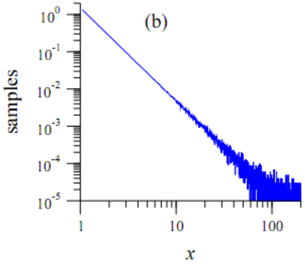
- straight line on log-log plot:

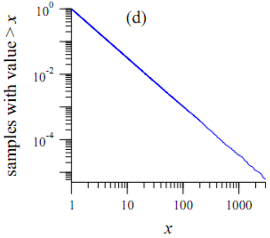
$$\mathbb{P}(X > x) \sim x^{-\alpha} \;\rightarrow\; \log(\mathbb{P}(X > x)) \sim -\alpha \log(x)$$

- plot cumulative distribution function rather than histogram

$$\mathbb{P}(X > x) \sim x^{-\alpha} \;\rightarrow\; \mathbb{P}(X = x) \sim x^{-(\alpha+1)}$$

- example from [Newman, 2004]





RuSSIR Russian Summer School in Information Retrieval

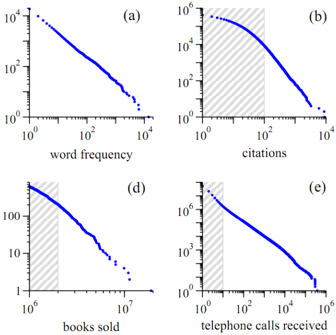
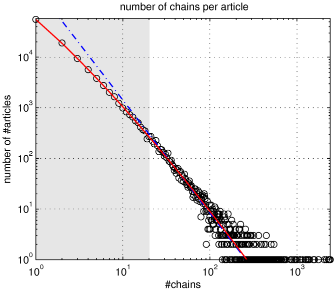
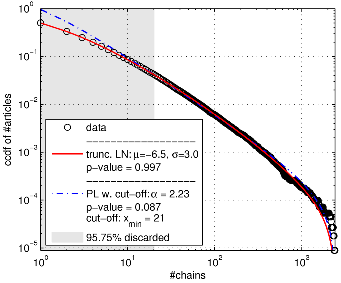


Figure: (a) Numbers of occurrences of words in the novel Moby Dick by Hermann Melville; (b) Numbers of citations to scientific papers published in 1981 until June 1997; (d) Numbers of copies of bestselling books sold in the US between 1895 and 1965; (e) Number of calls received by AT&T telephone customers in the US for a single day;

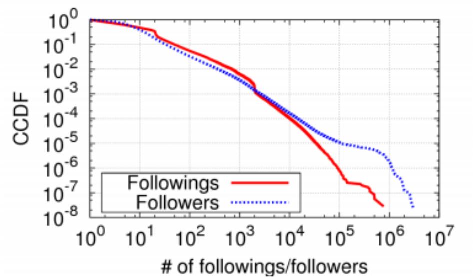
The number of discussion chains (A→B→A) per discussion page in Wikipedia [Laniado et al., 2011]





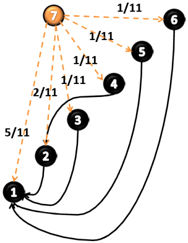
Power law
Log-log plot: examples

The number of followings (solid line) and that of followers (dotted line) on Twitter [Kwak et al., 2010].



- preferential attachment models: ‘rich gets richer’ approach
- directed and undirected versions [Barabasi and Albert, 1999][de Solla Price, 1976]
- growing network:
  - **time 1:** m nodes;
  - **time t:** add new node $[t+m]$ and link it to m old nodes;

$$\mathbb{P}([t+m] \to [i]) \sim \text{in-degree}([i]) + 1$$



- correlation coefficient:

$$\mathrm{corr}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]}{\sigma_X \sigma_Y},$$

  where σ_X and σ_Y standard deviations.
- if $\alpha_X,\ \alpha_Y \in (1, 2)$, then σ_X and σ_Y do not exist.

- *Angular Measure* [Resnick, 2007],[Volkovich et al., 2008]:
  - to measure extremal dependencies between power-law distributed parameters X and Y;
  - **rank transformation:**

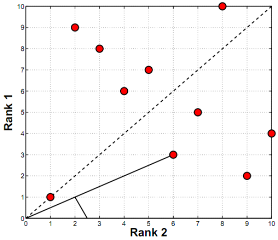
$$\{(X_j, Y_j), 1 \le j \le n\} \to \{(r_j^X, r_j^Y), 1 \le j \le n\},$$

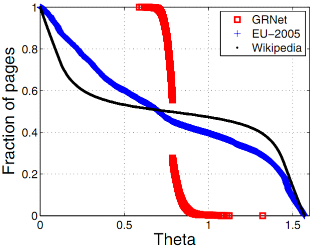
    where r_j^X and r_j^Y are the descending ranks of X_j in $(X_1, \ldots, X_n)$ and Y_j in $(Y_1, \ldots, Y_n)$ respectively.
  - **polar coordinate transformation:**

$$\text{POLAR}\left(\frac{k}{r_j^X}, \frac{k}{r_j^Y}\right) = (R_{j,k}, \Theta_{j,k}),$$

    where $\text{POLAR}(x, y) = \left(\sqrt{x^2 + y^2}, \arctan(y/x)\right)$

- empirical distribution of Θ for the k largest values of R:
  Dependence: measure is concentrated around $\pi/4$;
  Independence: measure is concentrated around 0 and $\pi/2$

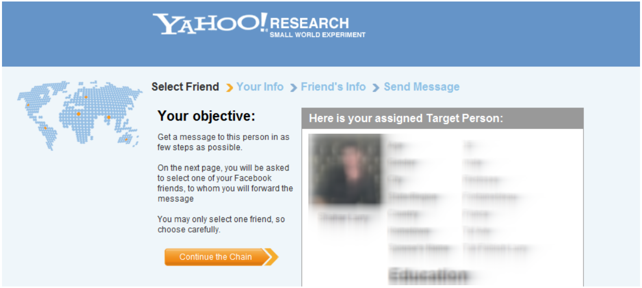




Diameter

- **diameter** is the “longest shortest path”
- **effective diameter** is the distance at which 90% of nodes can be reached.

- Many real graphs display small diameter
- '6 degrees of separation' [Travers and Milgram, 1969],[Dodds et al., 2003]
- `smallworld.sandbox.yahoo.com`



- Shrinking diameter [Leskovec et al., 2005].

Assortativity

- Mixing coefficient, or degree correlation, r allows to detect whether highly connected nodes preferentially link to other highly connected node [Newman, 2002]:

$$r = \frac{M^{-1}\sum_{e\in E} i_e j_e - \left(M^{-1}\sum_{e\in E}\frac{1}{2}(i_e + j_e)\right)^2}{M^{-1}\sum_{e\in E} i_e^2 j_e^2 - \left(M^{-1}\sum_{e\in E}\frac{1}{2}(i_e + j_e)\right)^2},$$

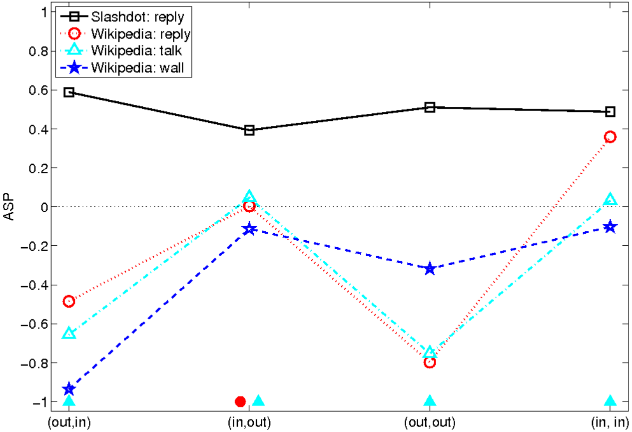
  where i_e and j_e are the degrees at the beginning and the end of edge e, E is the set of edges in the network and M its cardinality.
- **Assortative mixing** ($r > 0$) is present in many social networks;
- **Dissortative mixing** ($r < 0$) is present in food webs or in the Internet.

- Directed assortativity:
- Correlation between *in* and *out* degree of *source* and *target* nodes Foster et al. [2010]
- $(\alpha, \beta) \in \{in, out\} \rightarrow$ degree types of $(source, target)$

$$r(\alpha, \beta) = \frac{E^{-1} \sum_e [(i_e^\alpha - \bar{i^\alpha}) * (j_e^\beta - \bar{j^\beta})]}{\sigma^\alpha \sigma^\beta}$$

  - $E \rightarrow$ number of edges
  - $\bar{i^\alpha} = E^{-1} \sum_e i_e^\alpha$
  - $\sigma^\alpha = \sqrt{E^{-1} \sum (i_e^\alpha - \bar{i^\alpha})^2}$

Comparison of the directed Assortativity Significance Profile



Where ASP score is not significant ($|Z| < 2$), the corresponding ASP is marked with an appropriate symbol at the figure bottoms.

Influence propagation

Influence propagation:

- Spread of information (rumors);
- Model interest or trust;
- Innovation adoption;
- Expert finding;
- Social search and recommendations;
- **Viral marketing** (or "influence maximization"): Find a small subset of nodes in a social network that could maximize the spread of influences;
- etc.

- Add message ***"Get your free email at Hotmail"*** at the end of each sent email.
- jul. 1996: Hotmail.com launched
  aug. 1996: 20 000 subscribers
  dec. 1996: 100 000 subscribers
  jan. 1997: 1 million subscribers
  jul. 1998: 12 million subscribers

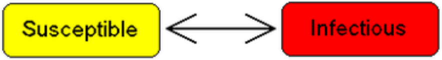
- Epidemiological models:
    - **SIR-model**: good model for Mumps;
    - **SIS-model**: good model for regular cold;

(**S** (for susceptible),**I** (for infectious) and **R** (for recovered))

- [Kempe et al., 2003] "Maximizing the spread of influence through a social network".
    - **IC** Independent Cascade model
    - **LT** Linear Threshold model

- initially: all nodes are in **susceptible (S)** state
  one node in the **infectious (I)** state;
- each time step
  1. **I** nodes attempt to infect their susceptible neighbors with probability β
  2. **I** nodes enter to the **recovered (R)** state (can not be infected again).

- all nodes are initially in *susceptible* (S) state, except for one node in the *infectious* (I) state;
- each time step
  (1) **I** nodes attempt to infect their susceptible neighbors with probability β
  (2) **I** nodes return to the susceptible state with probability λ or remain infected with probability $(1 - \lambda)$.



Independent Cascade (IC) model

- links have associated probability;
- when node v becomes active, it has a single chance of activating each of currently inactive neighbor w;
- the activation attempt succeeds with probability $p_{v,w}$.

Linear Threshold (LT) model

- node v has random threshold $\Theta_v \in [0, 1]$;
- node v is influenced by each neighbor w according to weight $b_{v,w}$ such that

$$\sum_{w \text{ is a neighbor of } v} b_{v,w} \leq 1$$

- node v becomes active when at least (weighted) Θ_v fraction of its neighbors are active

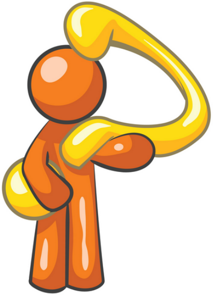
$$\sum_{w \text{ is a neighbor of } v} b_{v,w} \geq \Theta_v$$

Influence Maximization Problem:

- f(S) is ***influence*** of set of nodes S: the expected number of active nodes at the end of propagation, if set S is the initial active set.
- **Problem**: Given a parameter k (budget), find a k-nodes set S to maximize $f(S)$.
- NP-hard optimization problem for both IC and LT models;
- *Greedy Algorithm:* every round add node v^* into S such that v^* and S maximize the influence spread of f.

Algorithm 1 Greedy

Input: G, k, σ_m
Output: seed set S
1: $S \leftarrow \emptyset$
2: **while** $|S| < k$ **do**
3: select $u = \arg\max_{w \in V \setminus S}(\sigma_m(S \cup \{w\}) - \sigma_m(S))$
4: $S \leftarrow S \cup \{u\}$



Bibliography I

A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

D. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soci. Inform. Science*, 27(5): 292–306, 1976.

P. Dodds, R. Muhamad, and D. Watts. An experimental study of search in global social networks. *Science*, 301(5634):827, 2003.

J. G. Foster, D. V. Foster, P. Grassberger, and M. Paczuski. Edge direction and the structure of networks. *Proceedings of the National Academy of Sciences*, 107(24):10815–10820, 2010. URL `http://dx.doi.org/10.1073/pnas.0912671107`.

D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, New York, NY, USA, 2003. ACM.

Bibliography II

H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

A. N. Langville and C. D. Meyer. Deeper inside PageRank. *Internet Mathematics*, 1:2004, 2004.

D. Laniado, R. Tasso, Y. Volkovich, and A. Kaltenbrunner. When the wikipedians talk: network and tree structure of wikipedia discussion pages. 2011.

J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

M. E. J. Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89 (20):208701, Oct 2002. doi: 10.1103/PhysRevLett.89.208701.

Bibliography III

M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Arxiv preprint cond-mat/0412004*, 2004.

S. I. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*, volume 10. Springer Verlag, 2007.

J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

Y. Volkovich, N. Litvak, and B. Zwart. Measuring extremal dependencies in web graphs. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1113–1114. ACM, 2008. ISBN 978-1-60558-085-2.