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 $G = (V,E)$,
- $V$ is the set of vertices, or nodes,
- $E$ is the set of edges (edges may have weights)
Example

- ‘user ⇔ user’ graphs on the base of social interactions (e.g. friendship, communications: sharing, commenting)
Graph mining

Example

- \textit{‘user }\leftrightharpoons\textit{ user’ graphs on the base of social interactions (e.g. friendship, communications: sharing, commenting)}
Graph mining

Example

- ‘user ⇄ properties’ bipartite graphs
**Graph mining**


- 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)

![Image of Wikipedia article and discussion pages]
Graph mining

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.
Graph mining
Example: Wikipedia graphs intersection

- Jaccard coefficient of the overlap between the networks

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

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

<table>
<thead>
<tr>
<th></th>
<th>article-NW</th>
<th>talk-NW</th>
<th>wall-NW</th>
</tr>
</thead>
<tbody>
<tr>
<td>article-NW</td>
<td>1</td>
<td>0.11</td>
<td>0.09</td>
</tr>
<tr>
<td>talk-NW</td>
<td>0.11</td>
<td>1</td>
<td>0.35</td>
</tr>
<tr>
<td>wall-NW</td>
<td>0.09</td>
<td>0.35</td>
<td>1</td>
</tr>
</tbody>
</table>
Graph mining
Adjacency matrix

Matrix analysis
to represent a graph: Adjacency matrix

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

example from [Langville and Meyer, 2004]
local characteristics: in- and out-degrees, weighted degrees;
global characteristics: PageRank and modifications;

\[
PR(i) = c \sum_{j \rightarrow i} \frac{1}{d_j} PR(j) + \frac{1 - c}{N}.
\]

stationary distribution of an ‘easily-bored-surfer’ random walk on a graph
Power law
Graph mining

Quiz (1)

Average age distribution across social network sites
United States

Data source: Google Ads Planner (United States demographic data) www.pingdom.com

On average, how many hours a day do you use your iPad?

<table>
<thead>
<tr>
<th></th>
<th>Nov-10</th>
<th>May-11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 1 hr</td>
<td>15.20%</td>
<td>11.30%</td>
</tr>
<tr>
<td>1 to 2 hrs</td>
<td>36.65%</td>
<td>36.30%</td>
</tr>
<tr>
<td>2 to 5 hrs</td>
<td>38.29%</td>
<td>42.20%</td>
</tr>
<tr>
<td>5 to 8 hrs</td>
<td>7.30%</td>
<td>8.82%</td>
</tr>
<tr>
<td>More than 8 hrs</td>
<td>2.02%</td>
<td>2.30%</td>
</tr>
</tbody>
</table>
How Big Is Twitter, Really?

- 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.

Source: Twitter API
Graph mining
Power law

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 $\alpha$:

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

**Pareto principle:** for many events roughly 80% of the effects come from 20% of the causes;

$\alpha$ between 1 and 2: finite mean, infinite variance.
Power law

Log-log plot

- straight line on log-log plot:

\[ P(X > x) \sim x^{-\alpha} \rightarrow \log(P(X > x)) \sim -\alpha \log(x) \]

- plot cumulative distribution function rather than histogram

\[ P(X > x) \sim x^{-\alpha} \rightarrow P(X = x) \sim x^{-(\alpha + 1)} \]

- example from [Newman, 2004]
Power law
Log-log plot: examples

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;
Power law
Log-log plot: examples

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

\[
\mathbb{P}([t + m] \rightarrow [i]) \sim \text{in-degree}([i]) + 1
\]
correlation coefficient:

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

where \( \sigma_X \) and \( \sigma_Y \) standard deviations.

if \( \alpha_X, \alpha_Y \in (1, 2) \), then \( \sigma_X \) and \( \sigma_Y \) 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:

\[ \{(X_j, Y_j), 1 \leq j \leq n\} \rightarrow \{(r_j^X, r_j^Y), 1 \leq j \leq n\}, \]

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 $\Theta$ for the $k$ largest values of $R$:

- **Dependence:** measure is concentrated around $\pi/4$;
- **Independence:** measure is concentrated around 0 and $\pi/2$.
Graph mining

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]
- \((\alpha, \beta) \in \{\text{in}, \text{out}\} \rightarrow \text{degree types of (source, target)}\)

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

- \(E \rightarrow \text{number of edges}\)
- \(\bar{i}^\alpha = E^{-1} \sum_e i_e^\alpha\)
- \(\sigma^\alpha = \sqrt{E^{-1} \sum (i_e^\alpha - \bar{i}^\alpha)^2}\)
Where ASP score is not significant ( |Z| < 2), the corresponding ASP is marked with an appropriate symbol at the figure bottoms.
Influence propagation

Introduction
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
Influence propagation
Models

- 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
Influence propagation

SIR-models

- 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 $\beta$
  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 $\beta$
  2. I nodes return to the susceptible state with probability $\lambda$ or remain infected with probability $(1 - \lambda)$. 

**Diagram:**

- Susceptible
- Infectious
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) $\Theta_v$ 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* ← ∅

2. while |*S*| < *k* do

3. select *u* = arg max_{*w* ∈ *V* \ *S*} (σ_m(*S* ∪ {*w*}) − σ_m(*S*))

4. *S* ← *S* ∪ {*u*}

---


