Mining Large-Scale Network Data

Jure Leskovec (@jure)

Data mining, Statistics and Machine Learning have rich history and methods for analyzing ...

- ... tabular data
- ... textual data
- ... time series & streams
- ... market baskets

What about relations and dependencies?
Networks!
Networks
are a general language for reasoning about real-world systems
Human interactions
Brain
Media & Information
Infrastructure

Human cell

Economy
Internet
Network!
## Networks, why now?

<table>
<thead>
<tr>
<th>$10^9$ nodes</th>
<th>$10^8$ nodes</th>
<th>$10^7$ nodes</th>
<th>$10^6$ nodes</th>
<th>$10^5$ nodes</th>
<th>$10^4$ nodes</th>
<th>$10^3$ nodes</th>
<th>$10^2$ nodes</th>
<th>$10^1$ nodes</th>
<th>$10^0$ nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
<td>SOCIAL NETWORK</td>
</tr>
<tr>
<td>WWW</td>
<td>SOCIAL NETWORK SITES</td>
<td>POWER GRID</td>
<td>INTERNET</td>
<td>METABOLIC NETWORK</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Large-scale network data

- **Web is a sensor into humanity!**

- **Profound transformation in:**
  - How knowledge is produced and shared
  - How people interact and communicate
  - The scope of CS as a discipline
What do we do?

Networks + BigData analytics = Computational models of behavior

Finding Friends

- **Growing body of research captures dynamics of social networks**
  [Latanzi, Sivakumar ’08] [Zheleva, Sharara, Getoor ‘09] [Kumar, Novak, Tomkins ‘06] [Kossinets, Watts ‘06] [L., Kleinberg, Faloutsos ’05]

- **What links will occur next?** [LibenNowell, Kleinberg ‘03]
  - **Social network + Many user features:**
    - Location, School, Job, Hobbies, Interests, etc.
Friend Recommendation

- Learn to recommend friends
- Facebook link creation [Backstrom, L ‘11]
  - 92% of new friendships on FB are friend-of-a-friend
  - Triadic closure [Granovetter, ‘73]
- More common friends helps:
  - Social capital [Coleman, ’88]
Goal: Given a user, recommend friends

Idea: Learn PageRank scores
  - User features “guide” a random walk

FB Network

Set friendship strengths
  (strong edges to point towards future friends)

Run Personalized PageRank on a weighted graph

Recommend users with highest score

Results on Facebook Iceland:
- Correctly predicts 8 out of 20 (40%) new friends
- 2.3x improvement over previous FB-PYMK

Fraction of friending based on recommendations
Friend or Foe?

- Not just if you link to someone but also what do you think of them
- **Start with the intuition** [Heider ’46]
  - The friend of my friend is my friend
  - The enemy of enemy is my friend
  - The enemy of friend is my enemy
  - The friend of my enemy is my enemy

![Diagram showing balanced and unbalanced relationships](image-url)
Friend or Foe?

> 90% accuracy

Why organize friends?
- Filter and organize content
- Control privacy and access

All social networks have this feature:
- Facebook (groups), Twitter (lists), G+ (circles)
- But circles have to be created manually!
Discovering Social Circles

friends under the same advisor
CS department friends
college friends
‘alters’ \( v_i \)
‘ego’ \( u \)

family members
highschool friends

Discover circles and why they exist

Suppose we know all the circles

For a set of circles $c$, model edge prob.:
\[
(\ , \ ) \propto \left( - \cdot \left( \ , \right) \right)
\]
- $(\ , \ )$ ...is edge feature vector describing $(\ , \ )$
- ...circle parameters that we aim to estimate

Example:

\[
(\ , \ ) = \begin{bmatrix}
1 & \text{work : position : Cryptanalyst} \\
1 & \text{work : location : GC&CS} \\
0 & \text{work : location : Royal Navy} \\
1 & \text{education : name : Cambridge} \\
1 & \text{education : type : College} \\
0 & \text{education : name : Princeton} \\
0 & \text{education : type : Graduate School}
\end{bmatrix}
= \begin{bmatrix}
. \\
. \\
. \\
. \\
. \\
. \\
. \\
. 
\end{bmatrix}
\]
- How well do we recover human circles?
- Social circles of a particular person:
Media & Information
How does information interact with our personal social networks?

Information flows from a node to a node like an epidemic
- Since August 2008 we have been collecting 30M articles/day: 6B articles, 20TB of data
- Challenge:
  How to track information as it spreads?
**Goal:** Trace textual phrases that spread through many news articles

**Challenge 1:** Phrases mutate!

Mutations of a phrase about the **Higgs boson** particle.

Goal: Find mutational variants of phrases

- **Challenge 2**: 20TB of data!
- **Solution**: Incremental partitioning
the evidence shows beyond any doubt that the U.S. postal service pro cycling team ran the most sophisticated, professionalized and successful doping program that sport has ever seen.

I went to a number of women’s groups and said, ‘can you help us find folks?’ and they brought us whole binders full of women.

well, governor, we also have fewer horses and bayonets because the nature of our military’s changed. we have these things called aircraft carriers, where planes land on them.

with all due respect, that’s a bunch of malarkey … not a single thing he said was accurate.

this storm is dangerous and it’s critical to follow the advice of local emergency officials. if people are told to evacuate, they need to do it.

Visualization of 1 month of data from October 2012
MetroMap of “Israel”
- Observe times when nodes adopt the information

But where did the first node find the information?

Potential node-to-node spread

How did the information “jump”?

External Influence
Most infectious: Business, Entertain.

Least infectious: Travel, Art

Socially driven: Tech, Health, Entertain

Externally driven: Sports, Politics

Need few exposures: Travel, Tech

Need many exposures: Art, Science

Networks are a natural language for reasoning about problems spanning society, technology and information.
Conclusion & Reflections

- Only recently has large scale network data become available
- Opportunity for large scale analyses
- Benefits of working with massive data
  - Observe “invisible” patterns
Towards the Model of You

- Social networks — implicit for millennia — are being recorded in our information systems
- Software has a complete trace of your activities — and increasingly knows more about your behavior than you do
- Models based on algorithmic ideas will be crucial in understanding these developments
THANKS!

Data + Code:
http://snap.stanford.edu

@jure
References

- **Supervised Random Walks: Predicting and Recommending Links in Social Networks** by L. Backstrom, J. Leskovec. ACM International Conference on Web Search and Data Mining (WSDM), 2011.
- **Defining and Evaluating Network Communities based on Ground-truth** by J. Yang, J. Leskovec. IEEE International Conference On Data Mining (ICDM), 2012.
- **The Life and Death of Online Groups: Predicting Group Growth and Longevity** by S. Kairam, D. Wang, J. Leskovec. ACM International Conference on Web Search and Data Mining (WSDM), 2012.
References

- **Meme-tracking and the Dynamics of the News Cycle** by J. Leskovec, L. Backstrom, J. Kleinberg. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.
- **Structure and Dynamics of Information Pathways in Online Media** by M. Gomez-Rodriguez, J. Leskovec, B. Schoelkopf. *ACM International Conference on Web Search and Data Mining (WSDM)*, 2013.
- **Information Diffusion and External Influence in Networks** by S. Myers, C. Zhu, J. Leskovec. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2012.