CPSC 534L: Topics in Data Management – Social Networks

http://www.cs.ubc.ca/~laks/534l/cpsc534l.html

Laks V.S. Lakshmanan
Department of Computer Science
University of British Columbia
Who I am, what I do, where/when you can find me

- I do data management (modeling, query language design/optimization, logic + databases), data mining (relational, text, graph and other databases), data warehousing and OLAP, data cleaning, social/information networks, and recommender systems; more recently, fake news detection and bursting the so-called filter bubble.

- Office: ICICS 315.
- Email: laks@cs.ubc.ca; Phone: 2-3153.
Course Practical Matters

• **When & where we meet:** TR 11:00-12:30 pm, DMP 101.

• **Formal Prerequisites:** none. That said, a working knowledge of graphs, algorithms, basic theory, basic data mining/ML, basic DB will be an asset.

• We will use piazza for online discussion on course related matters. Sign up at piazza.com/ubc.ca/winterterm12017/cpsc534l
Course Objective

- Learn about interesting, useful and challenging problems involving social networks† (SN) and recommender systems (RS); more generally graph mining.
  - Modeling
  - Search/Recommendations
  - Communities, top-\(k\)
  - Marketing, Information/Influence Propagation
  - Algorithmic Issues
  - Analysis
- Focus on research.
- This is not an exhaustive list!

†As well as other kinds of networks.
Course Outline 1/3

• **Intro**: origins, early history, sociologist’s perspective, social capital, centrality, social web, web 2.0 and web 3.0, what we care about in this course.

• **Data Mining over Networks**:
  ◦ Link analysis & prediction. Team Formation.
  ◦ Social search. Event Detection from Social Media.
  ◦ k-core and k-truss. k-core and k-truss decomposition.
  ◦ k-truss communities.
Course Outline 2/3

- **Viral Marketing**: Influence & Information Propagation.

- **Recommender Systems**:
  - Content-based vs. **Collaborative Filtering**
    - Memory-based vs. Model-based.
    - Recommendations of novel objects.
  - Strategic Recommendations.
  - SN & RecSys.

- **Top-k query/search**.

- **Regrets**: An alternative to top-k.
Course Outline 3/3

• Opinion Mining & Social Networks (time permitting).

• Your Talks (interspersed with lectures, per topic).
  ◦ Bibliography to choose papers for topics will be constantly updated as we progress through the course.
  ◦ You’re welcome to bring your own idea for a paper presentation: I just need you to run it by me to make sure it’s appropriate.
Marking Scheme

- Homework Assignments: 20%.
- Class Participation: 5%
- Paper Presentation/Discussion: 30%
- Course Project: 45%

- What each of these means for you.
- Some topics – assigned reading.
- Questions?
Plagiarism

- Take the time to read it and understand it.
- Be sure to attribute everything that is not your own original idea/contribution, to the source you got it from.
- When in doubt, ask me.
Resources

- No text book, but there are excellent books:
Resources

- Research Literature (WWW, KDD, ICDM, SDM, ICML, WSDM, VLDB, SIGMOD, ICDE, NIPS, SIGIR, RecSys, selected journals, …).
- Slides.
- Discussion Report on Papers (for student talks).
- “Social Network” Reading Group: all of you are invited! Send me mail to sign up.
Tentative Schedule

- 1: Intro, DM/N start.
- 2: DM/N.
- 3: DM/N, VM start
- 4: VM
- 5: VM.
- 6: RS.
- 7: RS, Top-K
- 8: Top-K.

Schedule may vary dynamically as we adjust to pace.

*Your talks will be interleaved with lectures.*
Intro – Social Networks are Here
Social Networks – Intro.

- SN didn’t quite start like that!
- What is a SN from a sociologist’s perspective?
- Conventional data is *tabular*:
  e.g., \( R(\text{Name, Age, Gender, Salary}) \).
- rows = “cases”, “actors”, “subjects”, “observations” (tuples/users) and
- columns = “variables”, “measures” (attributes).
Social Networks – Intro.

- SN data: rows & columns = users*;
  - *groups/communities become nodes too.
- cell = relationship (tie) between users. (possibly weighted and/or labeled). → Matrix/Graph.
- nodes/vertices = “users”;
- edges (links) = relationships.
- attributes (properties) may qualify nodes and/or edges.
Social Networks – Intro.

- a network may be *homogeneous*: e.g., collaboration graph between scientists; facebook, LinkedIn networks; citation networks, some biological networks (protein-protein interaction).
Social Networks – Intro.

• Or it may be heterogeneous: e.g., publication datasets – pubmed, DBLP, etc.; flickr; last.fm; flixster; IMDB; any recommender system; ...
One more heterogeneous example
SN – Intro: some questions

- Simple analysis: which subjects (equiv., variables) are similar?
  - Interestingly, such analysis is frequently employed in RS (Collaborative Filtering!) [Users x Items].
- Consider a SN, i.e., a graph of users: similarity between Jack and Jill in terms of their friends.
- Similarity in terms of whom you like versus who likes you, in the case of directed graphs.
- Some common measures of similarity: cosine, Jaccard, (Pearson) correlation.
- Notion of local network: friends of user $u$ within $h$ hops; users in a community/group; in a school/class; in an income group; n/w need not be explicitly declared by users.
SN – Intro: some characteristics

- SN often exhibit a power law in a number of ways:

SN – Intro: some characteristics

- **Power law**: small #users have a large in-degree; most users have a very small in-degree (observe the long tail).
- When joining the network, users are more likely to connect to popular nodes.
- Most of the blogs are posted by a small #users.
- Most ratings/reviews of movies/songs come from a small #users.
- #downloads of songs; #citations of papers; populations of cities; #copies of genes in genomes; …
SN – Intro: some characteristics

- SN for us is just a graph $G = (V, E)$.
  - Nodes = users/actors/individuals/orgs/entities.
    - can be heterogeneous.
  - Edges = ties/relationships; can be of several types and be complex; can model using labels; can be directed.

- Social Network Analysis:
  - How do rumors spread (or innovations happen)?
  - How do diseases spread?
  - What is the avg degree of separation of the n/w?
  - Early experiments by Millgram.
SN – Intro: some characteristics

- Millgram’s experiments:
  - Controversial version: study how willing normal people are to obey instructions of an authority when the authority instructs them to act against their own conscience.
  - More relevant to SN: measure average min. #ties connecting two random people in the US by asking people to forward a mail to their contacts and seeing how many hops it took the mail to reach a certain target starting from random sources. Precursor of modern-day small worlds experiment.
  - The so-called six degrees of separation: [http://www.youtube.com/watch?v=V2biPHBGm3c](http://www.youtube.com/watch?v=V2biPHBGm3c)
  - Leskovec & Horvitz: Microsoft IM – large version of Milgram experiment.
Social Capital: Oft-used concept; no apparent formal def. in the literature. Here are some example def’s.

Burt 1992: [Social capital] is the sum of the resources, actual or virtual, that accrue to an individual or a group by virtue of possessing a durable network of more or less institutionalized relationships of mutual acquaintance and recognition.
SN – Intro: social capital

- The central proposition of social capital theory is that networks of relationships constitute a valuable resource for the conduct of social affairs, providing their members with "the collectivity-owned capital, a 'credential' which entitles them to credit, in the various senses of the word (Bourdieu, 1986: 249). Much of this capital is embedded within networks of mutual acquaintance and recognition.

For our purposes here, we adopt the latter view and define social capital as the sum of the actual and potential resources embedded within, available through, and derived from the network of relationships possessed by an individual or social unit.

The fundamental proposition of social capital theory is that network ties provide access to resources. One of the central themes in the literature is that social capital constitutes a valuable source of information benefits.

Sirianni & Friedland, 1998
SN – Intro: social capital

- Social capital – property of a group as a whole versus property of an individual based on its position in the group.
  - E.g., a member may “bridge” the gap between different groups.

- Intra-group perspective versus interactions with the outside world perspective.
SN – Intro: centrality

- Notions of centrlality:
  - **(Bidirectional) degree** as an indication of info. flow or activity.
  - **Betweenness**: Boundary spanner between different clusters; how many (shortest) paths pass through me? Potential point of failure.
Closeness: how close am I to other nodes, on avg.?

E.g. Distance distributions:

<table>
<thead>
<tr>
<th>Node</th>
<th>Dist=1</th>
<th>Dist=2</th>
<th>Dist=3</th>
<th>Dist=4</th>
<th>Dist=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \text{closeness}(u) = \frac{\sum_{v \in G} \text{dist}(u,v)}{|G|}. \]

- closeness(1) = 18/10;  closeness(5) = 16/10;
- closeness(3) = 16/10;  closeness(6) = 17/10;
- closeness(7) = 27/10. Any surprises?
SN – Intro: centrality

• so far, local perspective: centrality of a node.
• From a global perspective …
• How vulnerable is a network to single points of failure?
  ◦ How many (central) nodes/links can fail before n/w is disconnected?
  ◦ global measure.
  ◦ How are the various central nodes situated in the n/w?
• Avg path length in the network.
Example Networks

- Citation graphs, collaboration graphs.
- SN of corporate members.
- Wikipedia collaboration graph.
- Road and subway networks
- WWW, Internet
- Water networks
- Romantic relationships between people
- Online (social) networks like MySpace, Facebook, Flickr, last.fm, Youtube, MSN IM, …, LinkedIn, …
Social Web

- How people socialize or interact throughout the www, or more generally, “online”.
- Facebook, myspace, twitter, yahoo!360 – user-centric; LinkedIn – professional.
- Flickr!, del.icio.us, photobucket, tripit.com – item-centric (photos, bookmarks, trip stuff).
- call-graphs of mobile phone networks.
- Sometimes used interchangeably with Web 2.0.
- Sometimes as a future n/w similar to WWW which links, not just docs, but people, orgs, concepts, and resources, make links persistent against changes, resolve trust/privacy issues, promoting greater interoperability.
“Suppose you could go online and make relevant connections with others from whom you are separated by one, two, or three degrees? Suppose that while working on a solar energy project in California, you could use such a system to find an engineer in Shanghai whose experience is directly relevant to your project? Could the Internet be used to establish networks of trust that cross traditional borders? Can the Internet be better at supporting the ability of citizens to self-organize and participate in civil society?“

Web 2.0

- Connect to users not just to docs.
  - Social Networking [Facebook].
- Collaboratively create content, not just view existing content passively.
  - Collaborative tagging (folksonomy) [del.icio.us], rating, reviewing [amazon], blogging, wikipedia.
- Share content [youtube].
- RSS Feeds (change subscription).
- Hosted services, mashups, ...
- Non-conventional access devices (mobile phones, game consoles, TV(?)).
And Web 3.0.

- Complex (i.e., composite) search: e.g., “find a good camping site within 500 km of Vancouver, near a lake with kayak rentals, making sure the rental is working in the season” OR “find a nice comedy in town and a good resto with spicy food nearby” OR “find a good tropical destination for summer vacation that I can do in 1 week, in under $4000, which is family friendly.”

- “What is the best way I can fly to Tahiti under the current weather conditions?”
More on Web 3.0

- Web 3.0 browser searches different Internet sites and integrates and organizes info. for you.
  - Research on heterogeneous DBs is relevant.
More about Web 3.0

- Vague searches: “where should I go for dinner with my friend visiting from Africa?”
- Your browser knows your profile based on past behavior and uses that to generate a recommendation.
- Web 3.0 feels like a gigantic DB of everything, including you, your habits, preferences, actions, behavior, ...
  - Scary prospects for privacy!
  - "We don't need you to type at all. We know where you are. We know where you've been. We can more or less know what you're thinking about," Eric Schmidt, CEO, Google.
More on Web 3.0

- Knowing your search context: the “tropical vacation” search could trigger related activities/things to do based on collaborative filtering, but taking constraints into account.
- Personalization: the answer you get, for a given search query, depends on who you are, profile-wise.
- Web 3.0 is supposed to allow anyone to create a mashup!
- Supposed to leverage RDF & ontologies. Ontology authoring and maintenance – too complex for most people. → lot of interest in discovering ontologies from tagging data (folksonomy).
- Pretend all the info. you need is on a gigantic structured DB. How would you make many of the above a reality today?
Our focus in the course

- Technology is great, but underlying principles, formal foundations, theory, algorithms, and analysis/design are more fun.
- Emphasis on creativity and research challenges.