CS 6604: Data Mining Large Networks and Time-series

B. Aditya Prakash

Lecture #1: Introduction
Networks are everywhere!

Facebook Network [2010]

Gene Regulatory Network [Decourty 2008]

Human Disease Network [Barabasi 2007]

The Internet [2005]
What else do they have in common?
High School Dating Network


Blue: Male
Pink: Female

Interesting observations?
The Internet

Skewed Degrees
Robustness
Karate Club Network

(a)
Dynamical Processes over networks are also everywhere!
Why do we care?

- Social collaboration
- Information Diffusion
- Viral Marketing
- Epidemiology and Public Health
- Cyber Security
- Human mobility
- Games and Virtual Worlds
- Ecology
Why do we care? (1: Epidemiology)

- Dynamical Processes over networks

*CDC data: Visualization of the first 35 tuberculosis (TB) patients and their 1039 contacts*
Why do we care? (1: Epidemiology)

- Dynamical Processes over networks

  - Each circle is a hospital
  - ~3000 hospitals
  - More than 30,000 patients transferred

Problem: Given $k$ units of disinfectant, whom to immunize?
Why do we care? (1: Epidemiology)

~6x fewer!

[US-MEDICARE NETWORK 2005]

CURRENT PRACTICE

OUR METHOD

Hospital-acquired inf. took 99K+ lives, cost $5B+ (all per year)
Why do we care? (2: Online Diffusion)

> 800m users, ~$1B revenue [WSJ 2010]

~100m active users

> 50m users
Why do we care? (2: Online Diffusion)

- Dynamical Processes over networks

Followers ➔ Celebrity

Social Media Marketing

Buy Versace™!
Why do we care?
(3: To change the world?)

- Dynamical Processes over networks

Social networks and Collaborative Action
High Impact – Multiple Settings

Q. How to squash **rumors** faster?

Q. How do **opinions** spread?

Q. How to **market** better?
Dynamical Processes
= (a lot of) Networks
+ (some) Time-Series
Research Theme – Public Health

ANALYSIS
Will an epidemic happen?

DATA
Modeling # patient transfers

POLICY/
ACTION
How to control out-breaks?
Research Theme – Social Media

DATA
Modeling Tweets spreading

ANALYSIS
# cascades in future?

POLICY/
ACTION
How to market better?

Prakash 2015
Networks and Time-Series

Theory & Algo.
Comp. Systems
ML & Stats.

Biology
Physics
Social Science
Econ.

Prakash 2015
CS 6604: DM Large Networks & Time-Series
COURSE LOGISTICS
Course Information

- **Instructor**
  B. Aditya Prakash, Torgersen Hall 3160 F, [badityap@cs.vt.edu](mailto:badityap@cs.vt.edu)
  - Office Hours: 10-11am Tuesdays
  - Include string **CS 6604** in subject

- **Class Meeting Time**
  Tuesdays, Thursdays, 11:00am-12:15pm, McBryde Hall 655
  (time table shows McB 133C, ignore!)

- **Syllabus: Models, Theory, Algorithms, and Applications for Graph and Time-Series Mining**
  - Special Focus on how they relate to Cascade/Propagation-like Processes
Course Information

- **Keeping in Touch**
  
  Course website
  
  [http://www.cs.vt.edu/~badityap/classes/cs6604-Fall15/](http://www.cs.vt.edu/~badityap/classes/cs6604-Fall15/)
  
  Updated regularly through the semester
  
  – Piazza link on the website

---

**Description**

Can we guess if someone is sick from her tweets? How do opinions get formed in online forums? Which people should we immunize, to prevent an epidemic as fast as possible? How do quickly warn out of a group? Graphs—also known as networks—are powerful tools for modeling processes and situations of interest in real-life social systems, cyber-security, epidemiology and biology. They are ubiquitous, from online social networks, gene regulatory networks, to router graphs.

This course will cover recent research on the analysis of large networks (both theoretical and empirical), algorithms behind network problems, and their practical applications in various diverse settings. Topics include diffusion and virus propagation in networks, community detection, anomaly and outbreak detection, time-sequence segmentation and connections with work in public health, social sciences and cyber security.

**Course Information**

- Instructor: Prakash Baditya
- Times: Tuesdays and Thursdays, 11:10am-12:35pm
- Webpage: [http://www.cs.vt.edu/~badityap/classes/cs6604-Fall15/](http://www.cs.vt.edu/~badityap/classes/cs6604-Fall15/)

**Textbooks and Resources**

- There is NO required textbook. Recommended reading:
  3. Chuan Aggarwal, *Data Mining*. Springer 2015 (free PDF available accessible through the VT network only). Book Website: [here](#).
  5. See other resources (papers to datasets), [code](#) [here](#).

**Announcements**

- Welcome to the first class on 9/6/15.

**Schedule (tentative)**

- Lecture slides and readings: [here](#).
  1. Introduction
  2. Graph: Definitions and Graph Model
  3. Small worlds and Power Laws
  4. Dynamics of Networks and Models
Textbook

- **NO required textbook**
- **Recommended**
  
  David Easley and Jon Kleinberg: Networks, Crowds and Markets: Reasoning about a highly connected world. Cambridge University Press. 2010

Web page for the book (with FREE PDF!)


- **Other (excellent) related books:**
  
  See the course webpage
Pre-reqs

(A) Should enjoy the course 😊
(B) Ability to read research papers
(C) Background in
   1. Algorithms
   2. Probability and Stats
   3. Linear Algebra (helps)
   4. Graph theory (helps)
(D) Graduate-level Programming Skills (i.e. ability to use unfamiliar software, picking up new languages, comfortable with at least one of Python/C/C++/Ruby/Java etc. (Matlab/R a plus))
Rough outline

- Part (A) Networks: Structure and Evolution
- Part (B) Processes and Dynamics: Focus on Propagation
- Part (C) Graph analytics: Hadoop, Anomaly detection
- Part (D) Time Series Overview: Mining and Forecasting
Course Grading: Tentative

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>One</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework</td>
<td>10%</td>
<td>Presenting a paper in class/grading</td>
</tr>
<tr>
<td>Presentation/Grading</td>
<td>15%</td>
<td>Presenting a paper in class/grading</td>
</tr>
<tr>
<td>Course project</td>
<td>60%</td>
<td>BIGGEST component</td>
</tr>
<tr>
<td>Class Participation</td>
<td>10%</td>
<td>!= attendance 😊</td>
</tr>
<tr>
<td>‘Scribe’</td>
<td>5%</td>
<td>3 paragraph reaction post on Piazza</td>
</tr>
</tbody>
</table>

Class participation necessary!

- To get at least 5%
  - Non-trivial response to ‘reaction’ posts [see next slide] at least 6 times
  - Remember Piazza gives post statistics 😊
A 3 paragraph post on piazza after each class
  – Due 10pm the day of the lecture
  – Summary (4 lines)
    • What was the main technical content?
    • How did it relate to other topics?
    • What were the main doubts raised by students?
  – Expanding (3 lines)
    • Describe least one non-trivial related aspect not covered in the lecture
  – Brainstorming (4 lines)
    • Any open interesting questions (e.g. which readings missed)? Extensions?
    • What possible applications could benefit from the proposed material?
Scribe: Reaction Post

- 1 scribe per lecture
  - Everyone would scribe one lecture
- Will start after a few lectures
- Readings will be posted on Friday for the following week.
- The schedule will be posted on Piazza soon
  - Essentially will follow class-order alphabetically as on Hokiespa
Project (Groups of 2-3)

- Has to be **substantial**
- Can be
  - **Experimental**: evaluation of different algorithms and models on an interesting dataset(s)
  - **Theoretical**: considers a model (can be novel), or an algorithm, or a metric and derives a rigorous result about it (e.g. tighter bounds, surprising properties etc)
  - **Extension**: an extension or improvement of a method or model covered in class to a different or more general setting (e.g. time-varying, distributed, anytime, scalable, etc.) and experiments that justify the new proposal
- Can **NOT** be just a survey
- (no double dipping: should not get double credit for same work)
Project

- **Deliverables**
  - Project Proposal 15%
  - Project Milestone Report 10%
  - Final Report 20%
  - Final Presentation/Poster in class (TBD) 15%

- Proposal should contain a detailed survey of 3-4 pages
  - 6-8 papers in a topic of your interest
    - Discussed in or related to the course
Project

- Deliverables
  - Project Proposal
  - Project Milestone Report
  - Final Report
  - Final Presentation/Poster in class

- Start EARLY!

Rough due Dates

- ~Oct. 6
- ~Nov. 5
- ~Dec. 2
- Dec 3/8
One Homework

- To be released on ~Sept 15
- Due on ~Sept. 29 (beginning of class)
- Theoretical + Hands-on

- Start EARLY!
Other...

- Class moves at a fast pace! I’ll assume this is the only (or one of two) graduate level classes this semester for you

- Graduate class---so no spoonfeeding!

- *Announcement*: NO class on Thursday Aug 27
WARM-UP AND BASICS
A Question

- How many of you think your friends have more friends than you? 😊

- A recent Facebook study
  - Examined all of FB’s users: 721 million people with 69 billion friendships.
    - about 10 percent of the world’s population!
  - Found that user’s friend count was less than the average friend count of his or her friends, 93 percent of the time.
  - Users had an average of 190 friends, while their friends averaged 635 friends of their own.
Possible Reasons?

- You are a loner?
- Your friends are extroverts?
- There are more extroverts than introverts in the world?
Example

Average number of friends?

Source: S. Strogatz, NYT 2012
Example

Average number of friends
= ( 1 + 3 + 2 + 2 ) / 4
= 2

Source: S. Strogatz, NYT 2012
Example

Average number of friends
\[= \frac{1 + 3 + 2 + 2}{4}\]
\[= 2\]

Source: S. Strogatz, NYT 2012
Average number of friends
= \frac{1 + 3 + 2 + 2}{4} = 2

Average number of friends of friends
= \frac{3 + 1 + 2 + 2 + 3 + 2 + 3 + 2}{8} = \frac{(1 \times 1) + (3 \times 3) + (2 \times 2) + (2 \times 2)}{8}

Source: S. Strogatz, NYT 2012
Example

Average number of friends
= (1 + 3 + 2 + 2) / 4
= 2

Average number of friends of friends
= (3 + 1 + 2 + 2 + 3 + 2 + 3 + 2) / 8
= ((1x1) + (3x3) + (2x2) + (2x2)) / 8
= 2.25!

Source: S. Strogatz, NYT 2012
Actually it is (almost) always true!

- Proof?
Actually it is (almost) always true!

- Proof?

\[ E[X] = \sum x_i / N \]
Actually it is (almost) always true!

- Proof?

\[ E[X] = \sum x_i / N \]

\[ \text{Var}[X] = E[(X - E[X])^2] \]

\[ = E[X^2] - E[X]^2 \]
Actually it is (almost) always true!

- Proof?

\[
E[X] = \sum x_i / N
\]

\[
Var[X] = E[(X - E[X])^2]
\]

\[
= E[X^2] - E[X]^2
\]

\[
\frac{E[X^2]}{E[X]} = E[X] + \frac{Var[X]}{E[X]}
\]
Actually it is (almost) always true!

- Proof?

\[
E[X] = \sum x_i / N
\]

\[
Var[X] = E[(X - E[X])^2]
\]

\[
= E[X^2] - E[X]^2
\]

\[
\frac{E[X^2]}{E[X]} = E[X] + \frac{Var[X]}{E[X]}
\]

Essentially, it is true if there is any spread in # of friends (non-zero variance)!
Implications

- Immunization
  - We will see later, acquaintance immunization
    - Immunize friend-of-friend

- Early warning of outbreaks
  - Again, monitor friends of friends

---

Social Network Sensors

Contagious process passes through two phases, one in which the number of infected individuals exponentially increases as the contagion spreads, and one in which incidence exponentially decreases as susceptible individuals become increasingly scarce. These dynamics can be modeled by a logistic function. Central individuals lie on more paths in a network compared to the average person in a population and are therefore more likely to be infected early by a contagion that randomly infects some individuals and then spreads from person to person within the network. This shifts the S-shaped logistic cumulative incidence function forward in time for central individuals compared to the population as a whole (left panel). It also shifts the peak infection rate forward (right panel).

---

Figure 1. Network Illustrating Structural Parameters.

This real network of 105 students shows variation in structural attributes and topological position. Each circle represents a person and each line represents a friendship tie. Nodes A and B have different "degree," a measure that indicates the number of ties. Nodes with higher degree also tend to exhibit higher "centrality" (node A with six friends is more central than B and C who both only have four friends). If contagions infect people at random at the beginning of an epidemic, central individuals are likely to be infected sooner because they lie a shorter number of steps (on average) from all other individuals in the network. Finally, although nodes B and C have the same degree, they differ in "transitivity" (the probability that any two of one's friends are friends with each other). Node B exhibits high transitivity with many friends that know one another. In contrast, node C's friends are not connected to one another and therefore they offer more independent possibilities for becoming infected earlier in the epidemic.

Figure 2. Theoretical expectations of differences in contagion between central individuals and the population as a whole.

A contagious process passes through two phases, one in which the number of infected individuals exponentially increases as the contagion spreads, and one in which incidence exponentially decreases as susceptible individuals become increasingly scarce. These dynamics can be modeled by a logistic function. Central individuals lie on more paths in a network compared to the average person in a population and are therefore more likely to be infected early by a contagion that randomly infects some individuals and then spreads from person to person within the network. This shifts the S-shaped logistic cumulative incidence function forward in time for central individuals compared to the population as a whole (left panel). It also shifts the peak infection rate forward.

---

doi:10.1371/journal.pone.0012948.g001

doi:10.1371/journal.pone.0012948.g002
Network structure is important!!

- A network is a collections of nodes with relations between some nodes

Object: nodes, vertices $N$
Relations: links, edges $E$
System: graphs, networks $G(N, E)$
Networks and Graphs

- Networks: typically a real system
  - Metabolic Network, Social Network, etc.
- Graphs: typically the mathematical representation
  - Web graph, Planar graphs etc.

But we use it interchangeably
Networks: Which representation?

- Connect people who work together: professional network
- Connect authors and papers: co-authorship network
- Connect all people whose name is John Smith?
Networks: Which representation?

- Choice is important
  - In some cases there is a unique unambiguous representation
  - In most others, YOU have to choose
    - Depends on what you want to ask
Undirected vs Directed Graphs

- Undirected
  - Links are symmetrical
  - Examples
    - Friendships (on FB!)
    - Collaborators
Undirected vs Directed Graphs

- **Directed**
  - Links are directed
  - Examples
    - Following on Twitter
    - Phone calls
Graph connectivity

- Connected (undirected) graphs
  - There is a path between any two vertices
Graph connectivity

- Connected (undirected) graphs
  - There is a path between any two vertices
Graph connectivity

- Extension to directed graphs
  - **Strongly** connected: has a path from every node to every other node
  - **Weakly** connected: is connected if edge directions are disregarded

Why?
Next Class

- The Web as a graph
- More network properties
- The $G(n,p)$ model