NetSci High 2014
www.bu.edu/networks/workshop

Tuesday July 22 - Student Workshop Day 6

7:00 am – Breakfast (700 Commonwealth Avenue, Warren Towers, BU)

9:00 am – SCI 113

- "Talk Show #7"

9:20 am – SCI 128 & 130

- Parallel Workshop 1: How Google Works
- Parallel Workshop 2: Networks and Diseases

10:20 am – SCI 128 & 130

- Swap Workshops

11:30 am – Lunch (700 Commonwealth Avenue, Warren Towers, BU)

1:00 pm – SCI 113

- "Talk Show #8"

1:15 pm – SCI 113

- Informal Chat #3: Brian Keegan (Northeastern University)

2:15 pm – SCI 352

- Network Science Core Concepts in Small Groups

5:00 pm – Dinner (700 Commonwealth Avenue, Warren Towers, BU)

6:15 pm – SCI 352

- Network Science Core Concepts in Small Groups - *Returning Students Only*

Brian Keegan

I am a post-doctoral research fellow in computational social science with David Lazer at Northeastern University. My research there examines how social media like Twitter and Wikipedia can be used to improve predictive models of electoral success as well as performing small group experiments using Facebook.

I defended my Ph.D. in the Meida, Technology, and Society program at Northwestern University's School of Communication in September 2012. My dissertation examined the dynamic networks and novel roles which support Wikipedia's rapid coverage of breaking news events like natural disasters, technological catastrophes, and political upheaval.

I use methods in network analysis, multilevel statistics, simulation, and content analysis. My research employs a variety of large-scale behavioral data sets such as Wikipedia article revision histories, massively-multiplayer online game behavioral logs, and user interactions in a crowd-sourced T-shirt design community. I am interested in developing statistical models to understand the structure and dynamics of complex networks and activity bursts in online communities as well as how network structure influence organizational behavior.

I was born in Oregon and grew up outside Las Vegas, Nevada. I attended the Massachusetts Institute of Technology and received bachelors degrees in Mechanical Engineering and Science, Technology, and Society in 2006. I currently live in Cambridge, Massachusetts with my wife who is a graduate student at the MIT Media Lab.

For more information, visit: http://www.brianckeegan.com

How Google Works

This module, which was designed by Mariano Beguerisse-Dıaz, Sang Hoon Lee, and Lucas Leub, aims to introduce Google's PageRank algorithm for ranking pages on the World Wide Web.

To start the module, we ask the students to imagine a world without Google or other search engines and to develop their own strategies for finding information on the Web. Usually, one of the first ideas is to compile an exhaustive list of every Web page. We use this to introduce the idea of a crawler to navigate Web pages, and this leads naturally to the notion of representing the Web as a network with directed edges (the hyperlinks) between nodes (the Web pages).

We ask the students to think about how to figure out whether a Web page is relevant for the information one seeks, and this leads almost immediately to the issue of how one should rank Web pages in order of importance. One possibility that the students quickly bring up is that one can develop rankings based on the textual content of a page. (In one case, we had a good discussion about how we would try to use an automated method to distinguish the Amazon rain forest from Amazon.com.) We let the students know that the first Web search engines used to be

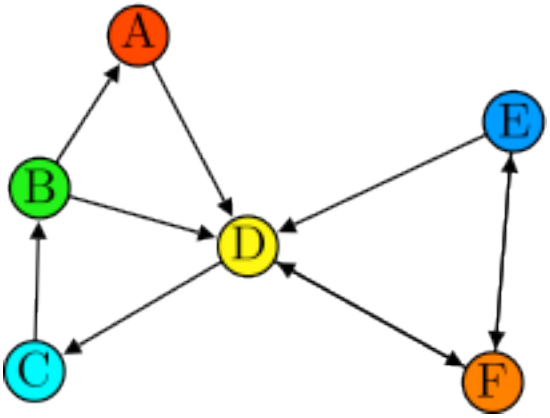

FIG. 2: An example of a directed network whose nodes we ask students to rank in order of importance. This network is strongly connected (so any Markov chain on it is ergodic), so we can ignore the problem of dead-end nodes (i.e., nodes with an out-degree of 0). However, we did discuss the notion of dead ends on many occasions that we ran this module. The ranking of the nodes in this graph from largest to smallest PageRank score (in parenthesis) is as follows: D (0.3077), F (0.2051), B (0.1538), C (0.1538), E (0.1026), and A (0.0769).

"curated" by hand, which limited how much of the Web could be explored. This limitation was an incentive to people to seek algorithmic methods to rank pages, which is what we want the students to explore. We ask the students to develop ideas for how to use the network structure of the Web to rank pages. (The difficulty of this transition in the discussion varied strongly from one to school to another.) Most of the time, the first structure-based ranking that the students propose is to rank Web pages according to the

number of incoming hyperlinks (i.e., according to in-degree). We discuss whether someone can cheat this system (as well as simple text-based systems) to improve the ranking of a page and whether better methods are available.

In Fig. 2, we show an example network that we use to help guide our discussion of how to rank Web pages. This example is particularly useful for moving beyond ideas for ranking based on the text on a page, as we can pretend that no such text exists (or that it is otherwise impossible to distinguish Web pages based on their text). We give the students a handout with this network (see the SOM), and we ask them to rank the nodes in order of importance (and also to indicate how they have defined "important").

Motivated by the fact that people often seem to explore the Web by "randomly" following hyperlinks—who hasn't done this on Wikipedia?—we ask the students whether they can develop a ranking method based on this idea. We use phrasing along the following lines: If we have a large number of monkeys—it is very compelling to refer to random walkers as "monkeys" [15]—who are clicking on hyperlinks randomly, what ranking would we obtain based on the number of times each page is visited. It can also be useful to discuss why Wikipedia is a "monkey trap", in the sense that many Wikipedia pages have high rankings in Google searches. (We occasionally discussed having one random walker versus having a large number of random walkers.) Using these questions, we introduce the rationale behind PageRank. Crucially, we try to avoid words like "eigenvalue" and "eigenvector" (and "ergodic", "Markov", etc.), though we do attempt to get the students to compute (by hand) the PageRank eigenvector for a network like the one in Fig. 2. They just don't know that what they are computing is called an eigenvector.

The example network in Fig. 2 is very instructive. Different choices for how to measure node importance (e.g., in-degree versus out-degree) lead to different rankings, and we have interesting discussions regarding which ranking is "correct". (These ideas could also be used to develop a module that focuses on centralities more generally—e.g., intuition related to the notion of "betweenness" sometimes comes up in Modules 2 and 3—as well as how one might change the notion of importance depending on the question one wants to answer.) An interesting feature of the network in Fig. 2, which is worth asking the students to try to prove, is that nodes B and C have the same PageRank score.

To compute the rankings, the students count the number of times the nodes are visited on different walks through the network. We and the students use two primary techniques for this calculation: (1) start from a uniform distribution and iteratively count the number of walkers on each node, or (2) try to identify relative orderings for the ranking of different nodes without calculating individual probabilities. The maximum out-degree in the example network is 2, which makes it easy to simulate a random walk by flipping a coin to decide which edge to follow. The students soon realize that one can get to node D from almost everywhere, and that it is indeed the most visited node in a (conventional) random

walk. This then makes it the most important node in this ranking scheme. The students then

realize that the nodes receiving edges from it (C and F) must come next in the rankings, and they discuss how to break the tie. Students also note that C and B are always visited the exact same number of times (as long as we don't stop the walk before a monkey has had the chance to leave C). In some cases, we were able to get the students to calculate the actual percentage of visits for each node rather than only determining the rank order.

When there is enough time, we discuss that a monkey gets "trapped" on a Web page with an out-degree of 0. This problem can be illustrated by adding a dead-end node G to the network in Fig. 2. We ask the students to think about how they can change their ranking methodology to be able to deal with this situation. This gives the opportunity to discuss the idea of a random walk with "teleportation" (e.g., at each node, one follows an edge with a probability p or otherwise chooses some other node in the network via a random process). Once the dead-end node has been added, the graph is no longer "strongly connected", which we can use to illustrate that even a small perturbation of a network can change its properties in a fundamental manner.

As part of this module, we sometimes discuss clever scientific ways to use Google—such as trying to measure the similarity between two football players by examining how often they show up together in Google searches [17].

Courtesy of Mason Porter, University of Oxford

From "Teaching Network Science to Teenagers," Network Science, 25 February 2013

Networks and Disease

This module illustrates how thinking about networks arises naturally when one tries to understand how diseases spread and how to develop good strategies to contain them.

We start this module by asking students to discuss the main characteristics of an infectious disease and to think about how it might spread (e.g., what is the main difference between a non-infectious disease and an infectious one). We encourage the students to think about a fictional disease that can only spread by shaking hands and to discuss how it might spread in their school. This quickly leads to the notion of disease spread along social networks in schools. For this discussion, it can be useful to distinguish online and offline social networks (and also to distinguish between the spread of ideas and rumours versus diseases).

We also briefly discuss what other types of networks (e.g., transportation networks or trade networks) might be important for understanding, containing, and preventing diseases. We steer the discussion towards how different types of network topologies can affect disease spread and vaccination strategies. If it is too expensive to vaccinate everybody, then who should be vaccinated? Some students brought up node labels in this discussion—one rebellious student proposed that the youngest people should be vaccinated because they (supposedly) had the longest left to live—though our primary focus was on network topology and how it affects disease spread. In some sessions, students brought up air travel, which led to a discussion of how such travel has changed the ways diseases spread. Students also pointed out that some diseases could be spread by insects like mosquitos, and we used this opportunity to introduce bipartite networks and to explain how vaccination strategies differ in this situation. For example, fumigation can eliminate many mosquitoes (reducing the number of one type of node) and slow down the spread of a disease.

The largest portion of the module is a hands-on activity in which we distribute handouts with various example networks (see Figure) to the students and ask them to devise possible vaccination strategies in each case if they are only allowed to vaccinate three or fewer nodes. The students realized quickly that this question was much harder to answer for some network topologies than for others. This was an interesting point of discussion, as it allows the students to consider how one might develop a vaccination strategy in real networks, which are much more complicated. Moreover, given that one needs to think about the answer even if one knows network structure exactly, we can discuss how to develop strategies when some (or even a lot) of the network structure is not known. We ask the students what would they do if we only know a network has a particular structure (e.g., suppose that one knows that it was generated using a Barabasi-Albert mechanism) but do not know anything else. This question generated a lively discussion. A useful hint for many students is to ask what would happen if we choose a node at random and then ask him/her to choose a friend to vaccinate (rather than vaccinating the original node). We also sometimes discuss the time-ordering of contacts in social networks and how that can influence disease spread.

In addition to discussing diseases specifically, it can be useful to encourage the students to think about other contexts in which “vaccination” strategies might be useful. One key question is the difference between the spread of an idea and a disease, and one might also wish to discuss other dynamical processes on networks.

This module is particularly nice for illustrating that mathematics shows up in many situations that the students (and their teachers) did not previously consider to be mathematical. This occasionally came up in discussions of viable careers for people who study mathematics at the university level, and we highlighted that nowadays mathematicians work closely alongside health professionals.

Figure on next page: Example networks for which one can apply different vaccination strategies (see the SOM). (A) star network, (B) circular lattice, (C) regular circular lattice with four neighbors (D) Erdős-Renyi random graph, and (E): Barabasi-Albert network.

Guiding questions for Networks and Disease

Introduction – 5 min

What are the main characteristics of an infectious disease?

Encourage the students to think about a fictional disease that can only spread by shaking hands. How it might spread in their school?

What other types of networks (e.g., transportation networks or trade networks) might be important for understanding, containing, and preventing diseases?

Activity –

What is one strategy we use today to prevent the spread of disease? --→ Vacination

What if it were too expensive or there are not enough supplies to vaccinate everybody? Who should be vaccinated? (H1N1 virus a couple of years ago)

Think about some of the networks you have seen over the last week and think about how a disease might spread on some of those networks. (10 min)

Pass out the network worksheet. Have the students work with their neighbor and devise possible vaccination strategies for each network. They can only vaccinate 3 or fewer nodes on each network. (5 min)

Have volunteers come up and share their strategies for each topology. Why? (10 min)

What if we don't know the network structure? What would your vaccination strategy be then? What if I randomly select a node and ask him/her to choose a friend to vaccinate? (friendship paradox) (5min)

Computer exercise – 20 min

Open firefox and go to http://vax.herokuapp.com/

Go through the tour first and then play the game trying to stop the spread of disease through vaccination and quarantine.

How would you vaccinate in each case?