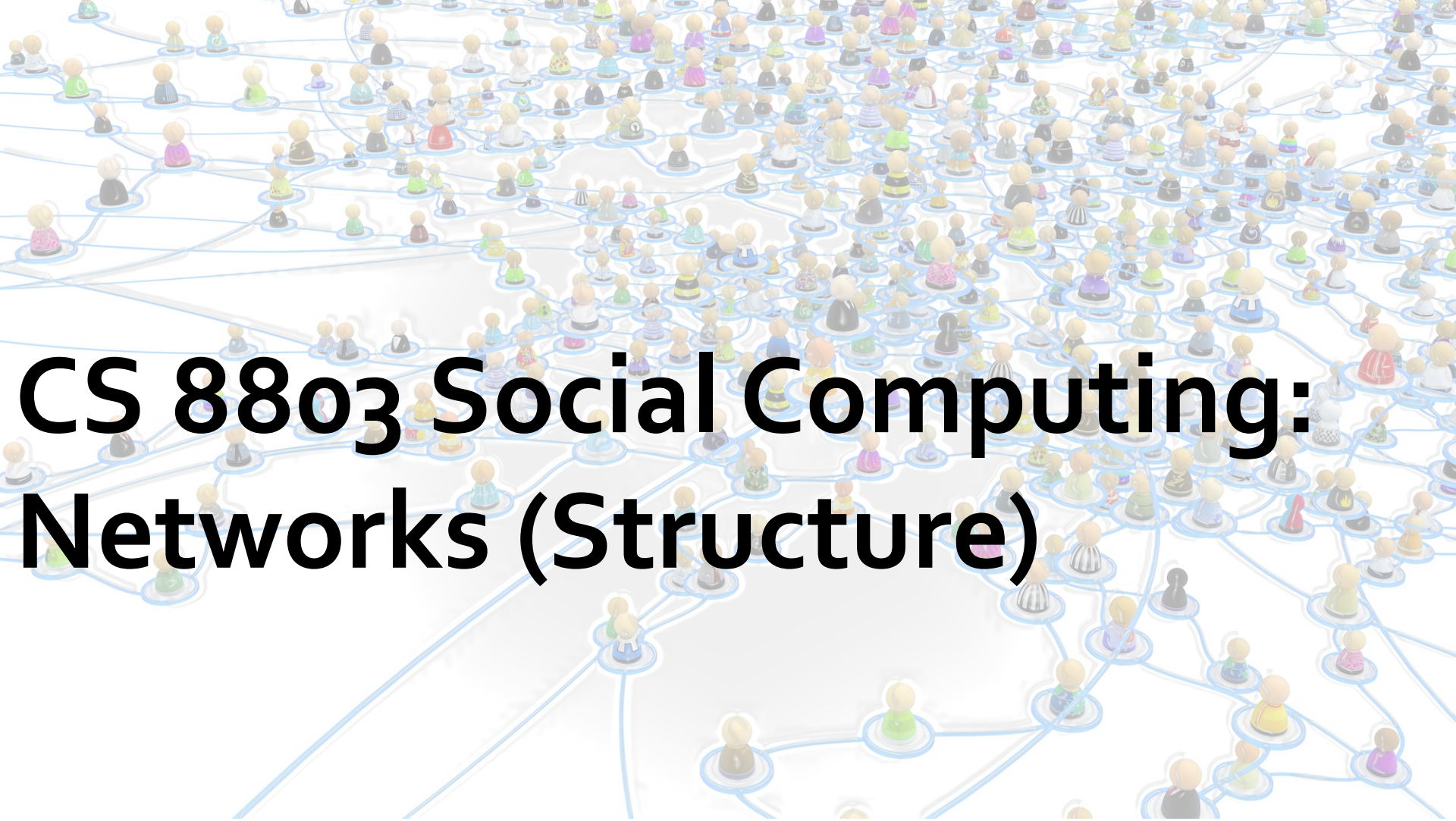


CS 8803 Social Computing: Networks (Structure)

Munmun De Choudhury

munmund@gatech.edu

Week 8 | October 6, 2014

Assignment II

- Build and evaluate a binary classifier that distinguishes between legitimate (ham) and spam SMS.
- 10% of class grade; 2% extra credit
- **Due: October 15, 3:30m**
- Deliverable: 3 page report
- Dataset to be downloaded from: www.munmund.net/courses/fall2014/notes/smsspamcollection.zip
- Action items:
  - Data preprocessing
  - Feature engineering
  - Building a classifier
  - Evaluation
  - Analysis of results

- Frigyes Karinthy in 1929 published a volume of short stories called "Everything is Different"
- He was the first proponent of the six degrees of separation concept, in his 1929 short story, Chains (Láncszemek)
- In his book the characters created a game out of the notion that "the world is shrinking":

> A fascinating game grew out of this discussion. One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth – anyone, anywhere at all. He bet us that, using no more than *five* individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances

An Experimental Study of the Small World Problem

Summary

- First sociological study of the "six degrees of separation"
- Empirically determine the maximum number of intermediaries it would require to reach anybody in the US
- Experiment conducted through forwarding of a set of snail mail letters, all targeted to a target in Massachusetts
- N=296 for two groups in Nebraska and Boston
- Main strategies involved in selecting the next point of forwarding: geographic and business
- Results:
  - 64 chains reached target
  - Completion rate was 31% for stock brokers in Nebraska whereas 24% for Boston
  - Funneling - Presence of a set of "hubs"/sociometric stars, through which most letters went through near the final target

Planetary Scale View on a Large Instant Messaging Network

Summary

- One of the first online studies of the small world phenomenon.
- Also the largest social network analyzed at that time: 180 million nodes and 1.3 billion edges.
- Instant messenger interactions showed stronger correlation over age and demographics.
- However interactions between opposite gender were more frequent and longer.
- Countries with historic and ethnic relationships showed stronger interaction patterns.
- Findings validate Milgram's theory: it is found that individuals in the network are on an average 6.6 hops apart.

Four Degrees of Separation

Lars Backstrom* Paolo Boldi† Marco Rosa† Johan Ugander* Sebastiano Vigna†

January 6, 2012

Abstract

Frigyes Karinthy, in his 1929 short story "Láncszemek" ("Chains") suggested that any two persons are distanced by at most six friendship links.[1] Stanley Milgram in his famous experiment [20, 23] challenged people to route postcards to a fixed recipient by passing them only through direct acquaintances. The average number of intermediaries on the path of the postcards lay between 4.4 and 5.7, depending on the sample of people chosen.

We report the results of the first world-scale social-network graph-distance computations, using the entire Facebook network of active users (≈ 721 million users, ≈ 69 billion friendship links). The average distance we observe is 4.74, corresponding to 3.74 intermediaries or "degrees of separation", showing that the world is even smaller than we expected, and prompting the title of this paper. More generally, we study the distance distribution of Facebook and of some interesting geographic subgraphs, looking also at their evolution over time.

The networks we are able to explore are almost two orders of magnitude larger than those analysed in the previous literature. We report detailed statistical metadata showing that our measurements (which rely on probabilistic algorithms) are very accurate.

1 Introduction

At the 20th World–Wide Web Conference, in Hyderabad, India, one of the authors (Sebastiano) presented a new tool for studying the distance distribution of very large graphs: HyperANF [3]. Building on previous graph compression [4] work and on the idea of diffusive computation pioneered in [21], the new tool made it possible to accurately study the distance distribution of graphs orders of magnitude larger than it was previously possible.

One of the goals in studying the distance distribution is the identification of interesting statistical parameters that can be used to tell proper social networks from other complex networks, such as web graphs. More generally, the distance distribution is one interesting *global* feature that makes it possible to reject probabilistic models even when they match local features such as the in-degree distribution.

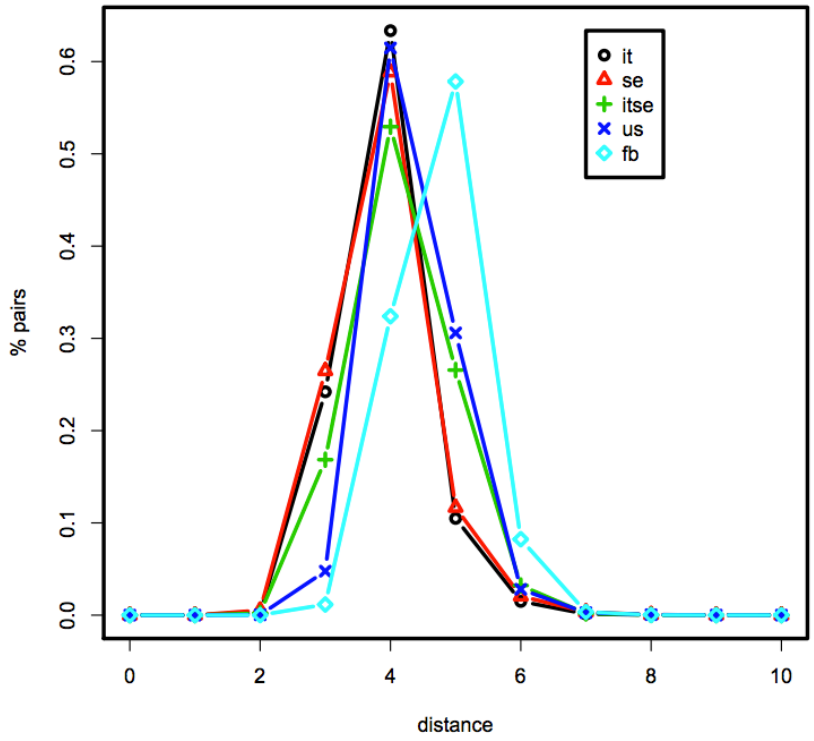
In particular, earlier work had shown that the *spid*[2], which measures the *dispersion* of the distance distribution, appeared to be smaller than 1 (underdispersion) for social networks, but larger than one (overdispersion) for web graphs [3]. Hence, during the talk, one of the main open questions was "What is the spid of Facebook?".

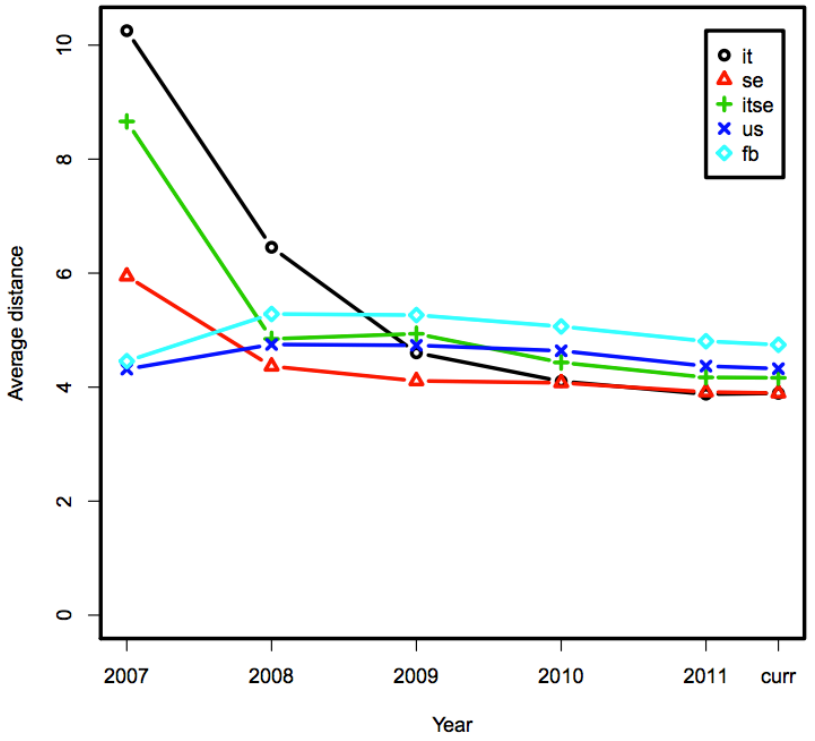
Lars Backstrom happened to listen to the talk, and suggested a collaboration studying the Facebook graph. This was of course an extremely intriguing possibility: beside testing the "spid hypothesis", computing the distance distribution of the Facebook graph would have been the largest Milgram-like [20] experiment ever performed, orders of magnitudes larger than previous attempts (during our experiments Facebook has ≈ 721 million active users and ≈ 69 billion friendship links).

This paper reports our findings in studying the distance distribution of the largest electronic social network ever created. That world is smaller than we thought: the average distance of the current Facebook graph is 4.74. Moreover, the spid of the graph is just 0.09, corroborating the conjecture [3]

*Facebook.

†DSI, Università degli Studi di Milano, Italy. Paolo Boldi, Marco


% pairs
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
distance
it
se
itse
us
fb


Average distance
0
2
4
6
8
10
2007
2008
2009
2010
2011
curr
Year
it
se
itse
us
fb

The Political Blogosphere and the 2004 U.S. Election: Divided They Blog

Summary

- First analysis of politics and elections on social media.
- 2004 Presidential elections were studied over blogs, particularly 4 A-list bloggers, over a two month period before elections.
  - 12470 posts from the left, and 10414 posts from the right
- Findings:
  - Conservatives and liberals were situated in contrastingly different and disconnected communities.
  - Difference was observed in terms of the news and other external content shared.
  - Conservative blogs were more tightly knit, in terms of links cited.
  - Liberals had stronger reciprocal connections.
  - Conservative blogs occasionally linked to liberal blogs whereas the reverse was not true.
  - Analysis of blog comments indicated stronger association within communities than between communities.

In Milgram's chain letter experiment, there was an assumption about a constant drop off rate as the letters traveled forward. What are the problems with this assumption?

Travers and Milgram had examined the incomplete chains. They assumed sometimes people didn't forward because they couldn't find an appropriate acquaintance. Could there be other reasons? E.g., you don't you RT everything you see on Twitter?

In Milgram's chain letter experiment, letter forwarding may imply a different notion of a friend compared to e.g., forwarding an email. Can these differences affect the number of hops?

Milgram did not after all investigate whether tie strength might play a role. How do you think tie strength would impact the so called “small world phenomenon”?

In Milgram's chain letter experiment, men were 10 times more likely to forward the letters than women. Why do you think it was the case?

The target was a stock broker in Massachusetts. And naturally the chains that were successful were either geographic or professional. How would you conduct the same experiment today to validate six degrees on personal social ties?

Leskovec and Horvitz found that 99.9% of the nodes in the graph of Live Messenger conversations were connected. Why do you think this was the case? Are Twitter or Facebook likely to be different?

Adamic and Glance only analyzed a handful of political bloggers. Would results differ for regular social media users?

Adamic and Glance only analyzed connections between conservatives and liberals. Could semantic analysis of blog content revealed something different?

Adamic and Glance only analyzed connections between conservatives and liberals. No consideration was made of the signed nature of ties. How would you use this concept on the political domain?

Next class

- Wednesday 10/8
- Topic: Networks (Time)
- There are assigned readings, due on Tuesday 11:59pm on Piazza.