Using the *Friendship Paradox* for Early Disease Detection

CS: 4980 Spring 2020

Computational Epidemiology

Tue, Apr 21

Social Network Sensors for Early Detection of Contagious Outbreaks

Nicholas A. Christakis[1,2*], **James H. Fowler**[3,4]

1 Faculty of Arts & Sciences, Harvard University, Boston, Massachusetts, United States of America, **2** Health Care Policy Department, Harvard Medical School, Boston, Massachusetts, United States of America, **3** School of Medicine, University of California San Diego, La Jolla, California, United States of America, **4** Division of Social Sciences, University of California San Diego, La Jolla, California, United States of America

Problem: Disease is assumed to spread stochastically on an *unknown* contact network of individuals. Which subset of individuals should we "monitor" in order to detect an outbreak early?

Question 1: The fact that the underlying contact network is unknown is key to their approach. What approach would you use if the network was fully known?

(Hints in the first three paragraphs of the Introduction)

Fully known network

If the network were fully known, we could compute *centrality measures*, e.g., *betweenness centrality.*

Definition (x_j, betweenness centrality of node j)

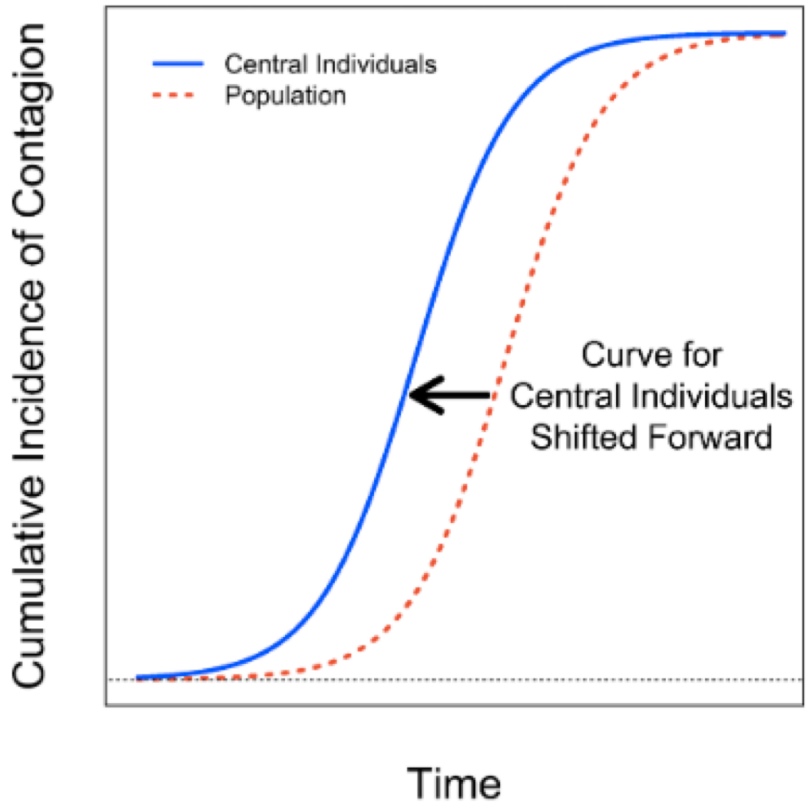
Let σ_{ik} denote the number of shortest paths between nodes i and k. Let σ_{ijk} denote the number of shortest paths between nodes i and k, that pass through node j. Then,

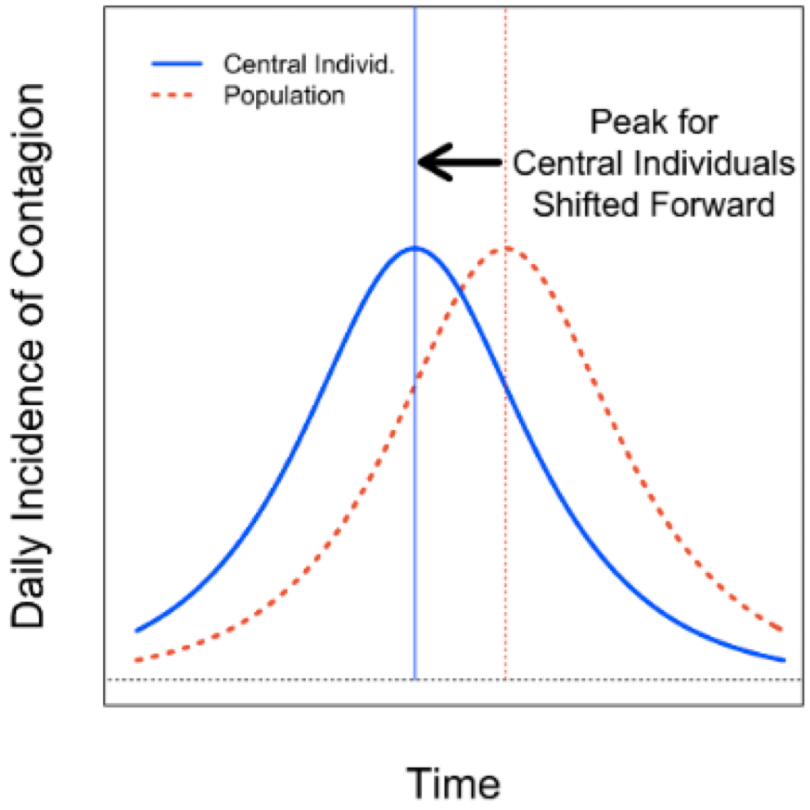
$$x_j = \sum_{i \neq j \neq k} \frac{\sigma_{ijk}}{\sigma_{ik}}$$

If we believe that nodes with higher betweenness centrality are more likely to be infected early, then we can pick the "top-k" nodes by betweenness centrality as sensor set.

Many other centrality measures

- Eigenvalue centrality, page rank centrality, closeness centrality, percolation centrality, etc.





Theoretical expectation of how these “central nodes” might behave as sensor sets.

Friendship Paradox

On average, your friends have more friends (on average) than you do!

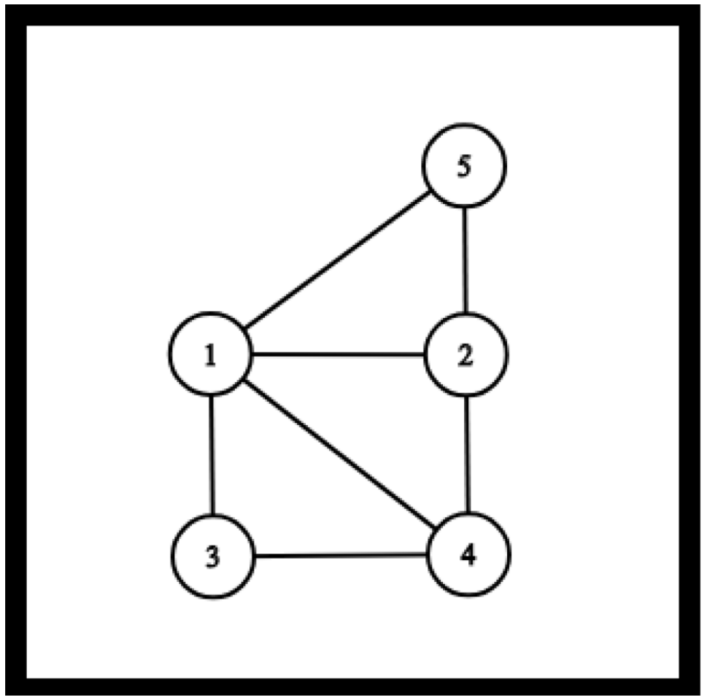
Under some circumstances, a “generalized friendship paradox” (e.g., your friends are happier than you or your friends are richer than you) holds!

(See https://arxiv.org/abs/1401.1458)

Question 2: (a) Mathematically, what is the friendship paradox claiming? (b) Is it true for all contact networks (graphs)? (c) Informally speaking, why is it true?

Example

1. Degree of a randomly chosen node = $\frac{4+3+2+3+2}{5} = \frac{14}{5} = 2.8$
2. Average degree of friends of a randomly chosen node
   - Node 1: $\frac{2+3+3+2}{4} = \frac{10}{4} = 2.5$
   - Node 2: $\frac{4+3+2}{3} = \frac{9}{3} = 3$
   - Node 3: $\frac{4+3}{2} = \frac{7}{2} = 3.5$
   - Node 4: $\frac{2+4+3}{3} = \frac{9}{3} = 3$
   - Node 5: $\frac{4+3}{2} = \frac{7}{2} = 3.5$



Final average = $\frac{2.5+3+3.5+3+3.5}{5} = 3.1$

Intuition

- High degree nodes have many neighbors (by definition!)
- A node chosen randomly is therefore likely to be a neighbor of a high degree node.
- Therefore, the average degree of its neighbors (friends) is likely to be higher than its degree.

This phenomena was discovered by sociologist Scott Feld and first described in this paper:

Feld, Scott L. (1991), "Why your friends have more friends than you do", *American Journal of Sociology*, **96** (6): 1464–1477

Generalized friendship paradox

- The friendship paradox was originally stated in terms of the *degree* attribute of nodes.

- It has now been shown to hold, both empirically and theoretically, for other attributes:
  - Your collaborators have published more papers than you have
  - Your twitter friends have spread more viral content than you have

- This paper shows that the friendship paradox holds for certain types of centrality attributes also (e.g., your friends are more central in the network than you are)

  Centrality-friendship paradoxes: when our friends are more important than us, Desmond J Higham, *Journal of Complex Networks*, Volume 7, Issue 4, August 2019, pages 515–528.

Sensor Set Selection

1. Pick a random subset of individuals (nodes).
2. Ask them to tell you who their friends are and use this set of friends as the set of nodes to monitor.

- The authors run this experiment on Harvard undergrads in the Fall of 2009. Read the details in Pages 2-3 and in the Supplementary document.
- They use the sensor set to monitor for influenza (seasonal or H1N1).

Question 3: Does the experiment, as implemented, match their intended algorithm? Can you identify any bias the sensor set might have that might be problematic for their results?

Experiment

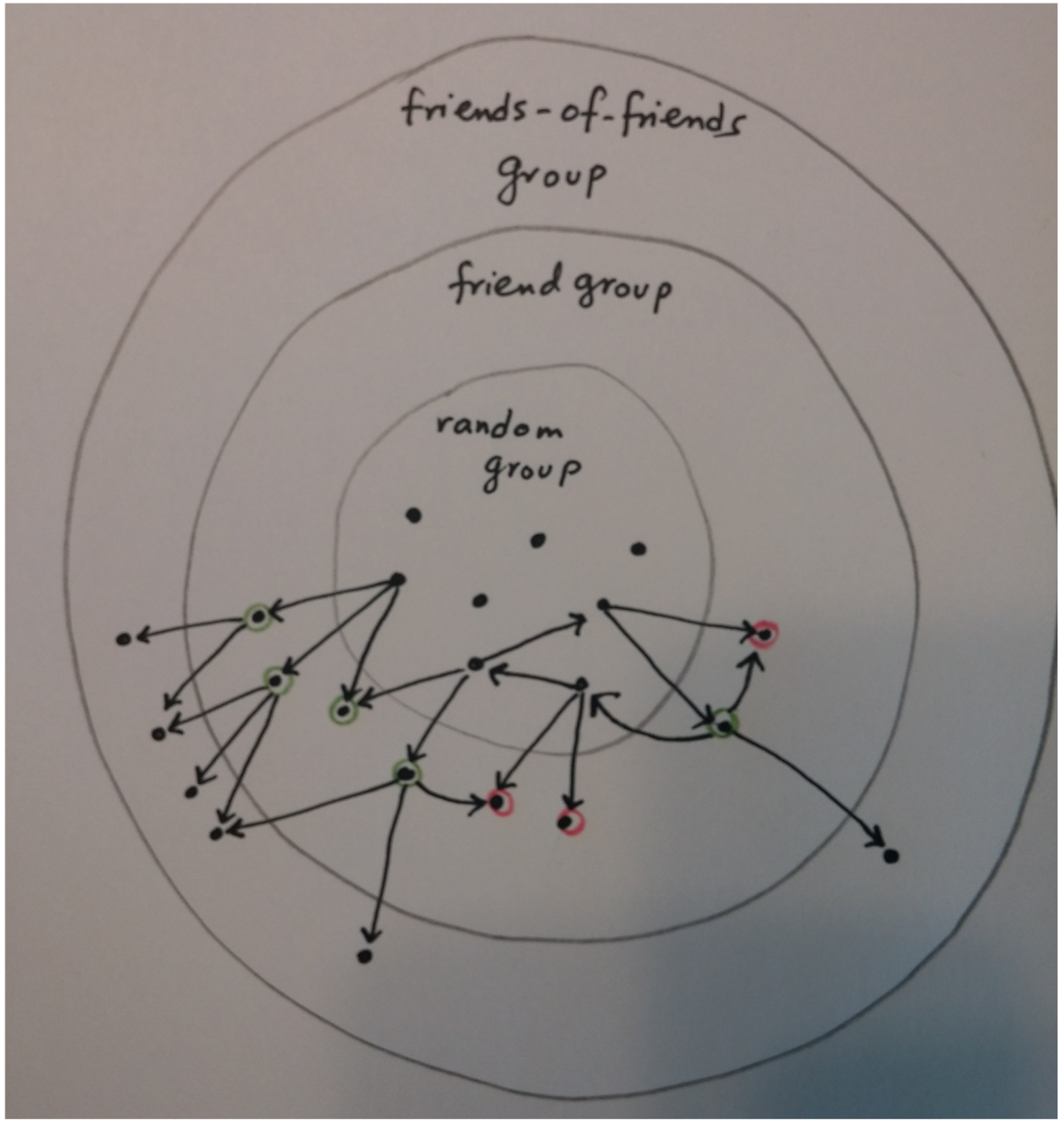
- Approached 1,300 randomly selected Harvard College students (out of 6,650) (starting on 10/23?)
- 396 (30%) agreed to participate and were asked to nominate up to three friends (random group).
- A total of 1,018 friends (not all unique) were nominated (average of 2.6 friends per nominator).
- Of the 950 unique friends, 425 (45%) agreed to participate (friend group).
- The random group and friend group has an intersection of 77. So size of experimental group = 396 + 425 – 77 = 744.
- The friend group (size = 425) nominated 1,180 of their own friends (average of 2.8 friends per nominator), yielding a friend-of-friends group of size 1,004.
- This friends-of-friends group had an intersection of 303 with the experimental group.

Induced graph

- Outdegree is at most 3, but indegree can be arbitrarily large.

- Nodes in friend group who have chosen not to participate have outdegree 0.

- Similarly, nodes in friends-of-friends group (who are not in other groups).



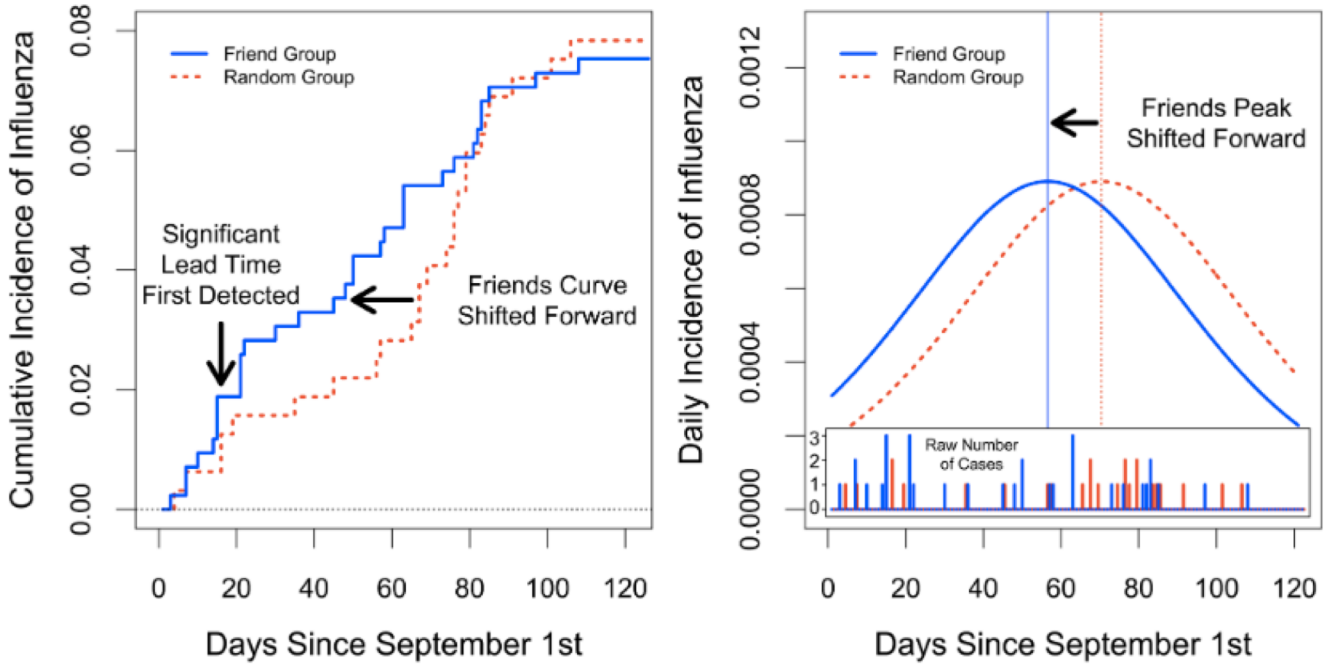
Reporting Influenza

- Tracked cases of formally diagnosed influenza among students in experimental group as recorded by University Health Services (Sep 1-Dec 31, 2009).

- Self-reported flu symptom email survey information (twice weekly: Mon, Thu), Oct 23-Dec 31, 2009.

- Self-reported Flu = fever of greater than 100F and at least two of the following symptoms: sore throat; cough; stuffy or runny nose; body aches; headache; chills;or fatigue.

Results: Early detection



This is Figure 3 (Page 4) and the most important figure in the paper. Understand what this figure is saying.

Question 4: What do you think of these results? Is the fact that the friends' set larger than the random set, a problem? These results are for diagnosis by medical staff; are the results for self-reported flu similar?

Results: Early Detection

- Friends curve for flu diagnosed by medical staff is shifted 13.9 days forward in time (95% C.I. 9.9–16.6).
- There is a smaller shift in self-reported flu symptoms (3.2 days, 95% C.I. 2.2–4.3).

Any thoughts on why the shift with self-reported symptoms is smaller?

Results: Graph Structure

Table S1: Summary Statistics for Friend Group and Random Group

	Friend Group		*Random Group*		*Mann-Whitney U*	
	Mean	*S.D.*	*Mean*	*S.D.*	*p*	*N*
Flu Diagnosis by Medical Staff	0.075	0.264	0.078	0.269	0.876	744
Self-Reported Flu Symptoms	0.325	0.469	0.310	0.463	0.678	744
In Degree	1.435	0.663	0.433	0.664	0.000	744
Out Degree	2.689	0.543	2.611	0.672	0.306	744
Betweenness Centrality (Percentile)	0.559	0.271	0.423	0.294	0.000	744
K-Coreness	1.673	0.553	1.414	0.565	0.000	744
Transitivity	0.142	0.231	0.148	0.274	0.039	721

This is a table from the supplementary material (Page 32). Understand what the graph-theoretic terms in the first column (e.g., Betweenness Centrality) mean.

Question 4: The `Friendship Paradox' is saying: your friends are more "central" to the network than you are. Is this borne out by the results in this table? Do you understand which graph these are computed on? Does the fact that these are computed on a partial graph, make these questionable?

Results: Graph Structure

- To compute betweenness centrality, k-coreness, and transitivity, the authors assume that every friendship is bidirectional.

- k-coreness
  - Repeatedly remove all nodes with one or fewer friends. Removed nodes are all assigned a value of $k = 1$.
  - Repeatedly remove all individuals who have two or fewer friends, giving them a value of $k = 2$.
  - …
- Higher value of k for a node implies node is more “central”.

- Transitivity $\equiv$ clustering coefficient

Confounding factors

The basic claim of the paper is this:

> Due to the “Friendship Paradox,” neighbors of randomly chosen nodes are more “central” in a network and therefore can better serve as an early detection sensor set, than a random sensor set.

Question 5: Has the paper shown this? Has the paper accounted for all the confounding factors you can think of? For example, could different vaccination rates in the two groups be responsible for what they’re observing?

	Friend Group		*Random Group*		*Mann-Whitney U*	
	Mean	*S.D.*	*Mean*	*S.D.*	*p*	*N*
Popularity Index	4.053	0.982	3.967	1.022	0.195	744
Self-Reported H1N1 Vaccine	0.200	0.400	0.188	0.391	0.685	744
H1N1 Vaccine at UHS	0.115	0.320	0.110	0.313	0.812	744
Self-Reported Seasonal Flu Vaccine	0.499	0.528	0.473	0.506	0.595	744
Seasonal Flu Vaccine at UHS	0.388	0.488	0.401	0.491	0.719	744
Female	0.720	0.450	0.627	0.484	0.007	744
Sophomore	0.176	0.382	0.235	0.425	0.049	744
Junior	0.259	0.439	0.238	0.427	0.522	744
Senior	0.322	0.468	0.276	0.448	0.172	744
Varsity Athlete	0.092	0.289	0.113	0.317	0.345	744

Thanks for your attention…

…any questions?