IFSC 4360/5360: Social Computing

Final Project Report

Northeastern University Instagram Followers: Network Graph of Interconnections

Tenzing Briggs

Objective:

The main goal of this project was to confirm that an Instagram account as huge as a major private university like Northeastern still is subject to trends relating to subcommunities that all follow the account. That is, the main goal was to see how interconnected subcommunities of Northeastern are. In other words, this goal was to visualize the Northeastern community through a network graph and discover what kinds of student communities appeared in such a graph and whether these communities were interconnected. Thus, another secondary goal was to become more familiar with web scraping software and with network graph visualization and analysis software.

Significance:

The applications of this project relate both to network analysis and to an understanding of Northeastern University. For the first, this project could be applied to show whether smaller samples of a larger sample (i.e., Northeastern has 117,000 followers but our dataset looks only at the most recent 807 followers) can still be used to effectively discover network characteristics. Another application would be to better understand Northeastern's social media imprint and makeup; it could be used by Northeastern marketers, recruiters, and officials to discover influential students or connections that span subcommunities and to see if Northeastern contains a "strength of weak connections" network.

Challenges:

The challenges faced in this project were how new I am to social computing methods (namely, programming in Python for API use) and the limit of how many nodes could feasibly be extracted from the Northeastern University Instagram account. Thus, a major challenge was finding an Instagram API web scraper that could be used with minimal programming or Python use. Another challenge was how to combine network graph software and close investigation of

profile characteristics; that is, because the nodes extracted were such a low number in comparison to the full follower number, a challenge was what to make of the small number of connections and how to visualize them.

Methodology:

The problem is how to discover communities, their kind, and their interconnection. Approaches to this problem would be topic modeling, sentiment analysis, toxicity analysis, or network analysis: all of these can identify communities of one or another. E.g., topic modeling can find communities based around shared topics and ideas, and sentiment and toxicity analysis can find communities based on shared outlook. Network analysis can identify communities based on relationships (edges) between students or followers (nodes).

I chose a network analysis approach because it can see if Northeastern University benefits from "strength of weak ties"—that is, does it have sparsely connected subcommunities (of strong ties) that can disperse information quickly throughout it? Another reason for choosing network analysis is the ability for it to show directed connections. Instagram analysis needs to consider the direction of relationships and connections, since Instagram friendship is directed (i.e. someone following another might not necessarily be followed back). Topic modeling, sentiment analysis, and toxicity analysis might find shared ideas, sentiments, and attitudes (respectively), but they might struggle to see where these things *originated* or *which* students brought such things to other students in a directed network like Instagram.

For the technical aspect of this approach, I used Twitter API and network visualization software: Phantombuster.com's Instagram Follower Collector API and the network graph software Gephi. Phantombuster's API was used to create a CSV list of the most recent 804 followers of the Northeastern University official Instagram account. Any private accounts were removed from the list (their followers are private). Then, this CSV was fed back into the Phantombuster API, creating a CSV of the first 500 followers of the 361 remaining who followed Northeastern, to see any interconnections between Northeastern followers (however, because I scraped twice, some nodes were overlapped, so that more than 500 of their followers were discovered). From this list, any Northeastern followers with 0 followers were deleted.

In both CSVs, the Excel Substitution formula removed the URL aspect of the extracted profiles, leaving just the username. Both CSVs were combined into one with, and all unnecessary information columns deleted (such as fullName, isPrivate, isVerfied, etc.) and a weight of 1 added to every row. This was then uploaded to Gephi and manipulated within the software for aspects like layout (OpenOrd), appearance (Modularity Class), and statistics (modularity and degree).

There were 2 node classes of nodes in our data: 1) the Northeastern followers, and 2) the followers of them. The majority of the class #2 were unconnected to all but the Northeastern

follower they were scraped from; few were **both** Northeastern followers and followers of other followers. Removing unnecessary nodes meant removing all nodes of **only** class #2 **that were not followers of 2 or more** of class #1; all nodes with a 0 in-degree **and** a 1 out-degree were deleted. After Gephi recalculated degree, all nodes with 0 in-degree **and** out-degree were deleted to remove any points. With the resulting network graph, notable profiles were analyzed qualitatively.

Analysis of Results/Findings:

After looking at the resulting network graph, several things became apparent. First, the coloring from modularity class indicated that there are, indeed, communities present, with a modularity score of 0.83. There also appear to be connections between these communities, shown by how some class #2 nodes follow 2 or more class #1 nodes across modularity groups. Influential students appear, such as maxgreenfield24, ryan_dosouza17, and jack_curtin16. Corporate entities also appear, and in many instances they share the same group of followers, such in the case of villagespeech and bostonbrowhoney. These corporate entities seem to connect to Northeastern by being made my people attending or living around Northeastern, like the example of forallthings digital (run by a student). Organizations also appear, like notyourtherapy.nu and northeasternneveragain, with names implying they are run by students or somehow Northeastern affiliated. Notably, few of the class #1 nodes (the main followers of Northeastern) directly follow other class #1 nodes, despite most class #1 nodes originally having hundreds of follower nodes previous to the deletion of 0 out-degree nodes. This implies that, with the hugeness of the Northeastern instagram, connections are primarily weak ties. Finally, hierarchies appear in a few places; e.g., princessrainlin, maxgreenfield24, and corum9991ig all are influential nodes which point from one to the other in a descending fashion.

Conclusion:

In terms of future directions, more research could use sentiment analysis or topic modeling to further understand the impact of the Northeastern follower makeup. For example, sentiment analysis could be used on organization or corporate followers to see if common sentiments appear across different communities. Topic modeling could then further identify student connections or common threads. Finally, this research could be furthered by more in-depth analysis of the student profiles in relation to other social media, such as using their account full name information to track down the students' current university and their exact class or year via more academic platforms like LinkedIn.