Instagram Following Relationships for Social Network Analysis

Kirin Danek
Kabir Hinduja
Seoeun Kim
Samuel Wang

Code available at https://github.com/swangtree/instagram_CN_graph

Introduction

Project Overview

This project investigates phenomena in a simple social network derived from instagram account and following-follower relations. We focused on scraping follower-following data from 13 accounts: 11 students and one instructor in the Spring 2025 Complex Networks course, as well as dis.copenhagen, the promotional account for the DIS Study Abroad in Scandinavia program based in Copenhagen, Denmark. The graph is created using the python NetworkX library; individual accounts are represented by nodes, follower-following relationships are represented by edges.

Motivation

Social Media platforms are a powerful tool of connection and socialization in current times and hold a central place in the contemporary digital landscape. These platforms and the social networks that exist there are not only digital representations of real life relationships but can also directly begin, maintain, transform, and influence social relations. Understanding the structure, shape, and behavior of social networks is critical in industries such as marketing, politics, and media. We wanted to gain a more nuanced understanding of social networks through conducting this project, at least on a very small, controlled scale. We also are personally invested in uncovering interesting patterns that exist within our own accounts and close ties, which is why we sampled Instagram data from the Complex Networks class.

Background

Social Networks can be modeled using graphs. Nodes represent individual accounts, and edges represent following-follower relations. In our directed graph, nodes will have outgoing edges to accounts they follow and incoming edges to accounts that follow them. In the undirected graph, accounts with any following/follower connection share an edge. By doing simple analysis on these graphs, we can measure and investigate influence, accessibility, and social roles of accounts in the network. We primarily used the NetworkX library in our project, as the library offers an expansive list of functions and can be easily used to analyze networks.

Research Questions

Research question 1: Sponsor and Influence Simulation

When trying to find a real world use case that leverages the network of influence we constructed, we wanted to simulate a topic with concrete data that would make the simulation more realistic, and a topic that had a unique connection to the class. The impromptu experience of attending a Livingston FC football match in Scotland was one of the highlights of our trip. All of these factors combined with the experience one group member had surrounding marketing with the head of FCK women, have led us to choose to simulate sports marketing through our complex network. Specifically, we will be looking at *which nodes would be the most effective for a sports team to sponsor if they wanted to grow their brand within our classes DIS students network*.

There were two additional factors which led to the construction of this question. While the network we will be working with is very large, it pales in comparison to the size of social media networks analyzed professionally, and contains many one degree nodes otherwise disconnected from the graph. This is why we choose to focus the question on our class's DIS student network, which we will be defining as a subset of our network consisting of any of the nodes with a degree of at least two, where our network is more useful. The relative smallness of our network, combined with the very low percentage impact individual social media interactions possess (measured in prior reports and discussed later on) led to the selection of a brand (sports teams) with very frequent social media activity and interaction opportunities. We will also still be making some favorable assumptions regarding statistics discussed later. These decisions give us a social media network which we can be both insightful and evolving.

Question 2: Connecting Accounts

Our social network reveals interesting social information about peers, celebrities, and other central figures to the Complex Networks Spring 2025 class. As a result, we are curious as to what connections classmates might have with each other apart from the class. We would love to, for instance re-create Stanley Milgram's small world experiment, but unfortunately this is out of our reach because our data only contains edges that represent an Instagram account following or being followed by a class member. This problem similarly limits our ability to compute centrality measures, which is a significant challenge. We attempt to bypass this issue by creating synthetic follower/following relationships among the accounts that are followed by or following class members, and constructing a three-partition graph in order to compute betweenness centrality on the subset of nodes representing class members when ignoring class-related connections.

Question 3: Asymmetric Cycles

Finally, moving beyond individual connections to local network structures, we investigated the smallest cycles within our graph – simple cycles of length three. In a directed network, these small cycles can reveal nuances about reciprocity and potential influence flows. A cycle where all connections are mutual (A follows B, B follows A, etc.) suggests a tightly-knit group, while asymmetric cycles (e.g., A follows B, B follows C, C follows A, but not all reverse connections exist) might indicate directed pathways or hierarchical relationships. We aimed to understand the prevalence of these symmetric versus asymmetric cycles within the network's core (the largest strongly connected component) and characterize individual nodes based on their participation in each type, shedding light on their local structural embedding and roles in reciprocal versus directed local interactions.

Data

Data Collection

The data was collected from Instagram over the course of a month from April 1st to May 1st. Since followers and following data cannot be collected for private Instagram accounts through the API, the only method for collecting Instagram follower data is through web scraping. Normally, follower and following data is not publicly available through viewing an Instagram

profile account, but it’s possible to view the list of an account's followers and following lists when viewing the account while logged into another account is mutually connected (both accounts follow each other). Everyone in the Complex Networks class (including the instructor) has an Instagram account, and most mutually follow each other, so data was able to be collected through manually scraping the profiles.

The process for data collection is as follows: First, the profile account was loaded on a web browser, then a javascript script was used to scroll to load all of a profile’s followers or following list. After the html content was loaded, the page content was scraped using BeautifulSoup to collect the username, name, profile url, and profile image url. For the purposes of this paper, only usernames were used to create the graph, but other node data could be used in future analysis. If User A follows User B, then a directed edge was created from User A to User B in the NetworkX graph. Data for 10 students in the class was fully collected, as well as for the instructor. Account data for Noah was incomplete, but a majority (789 out of 1,169) of his followers' data was included. Due to inherent limitations in how the data was collected, only 4,856 out of the approximately 10,400 followers for the dis.copenhagen account was able to be collected, although all of the following data was collected.

Data Description

The general structure of the graph is below in Figure 1. It can be viewed as a modified bow-tie structure, with a small, almost completely connected core of the class member of the Complex Networks class, with edges going into the DIS Copenhagen account node. The DIS account has a large number of nodes with edges leading into it, which are mostly students that have studied abroad or are interested in studying abroad. Like mentioned above, there are very few edges leading out from the DIS account. There’s also many accounts with mutual connections from class members, who are friends of class members. There’s also many public accounts that are followed by class members that have edges coming out from the strongly connected component, but not back in.

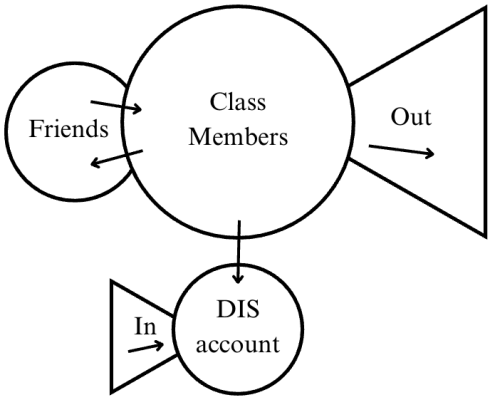

Figure 1: Basic data structure

Basic follower (In-degree) and following (Out-degree) counts for class members and dis.copenhagen are shown in Table 1.

Account	In-Degree	Out-Degree
swang330	685	1105
alex.kalis	1235	1675
alexandrapurdy_	711	869
kabir_aho	828	2346
kirindanek	825	1535
liam_hochman	817	1678
noahpurow	789	1603
seoeunki.m	1136	1403
will.deley	631	704
zach.annuik	604	1076
yota.katsikouli	147	185
dis.copenhagen	4857	240
elybrayboy	892	1471

Table 1: Follower (In-Degree) and Following (Out-Degree) counts for class members and related accounts.

Strongly Connected Component (SCC) in the Directed Graph

A strongly connected component is a subgraph of a graph where every vertex is reachable from every other vertex. We will use our directed graph to determine the size and shape of the strongly

connected component because if the graph were undirected, all the nodes would be in the SCC. There are 14,673 nodes not in the strongly connected component because they do not follow or are followed by any account in the main SCC.

Overall, 33% of the nodes in the directed graph are also in the SCC. This means our graph is highly fragmented since two-thirds of the nodes in our directed graph are in no way reachable from the nodes in the SCC. Since our graph is a social network, directionality and the connectivity observed here is relatively common. The 'dense' inner SCC is expected in a social network where not every connection will be in mutual directions.

Distribution of SCC sizes (Size:Count):

- Size 1: 14673 component(s)
- Size 7278: 1 component

Longest Shortest Path in Directed Graph

The diameter of the largest strongly connected component in our directed graph is 6, where the source of the path starts at divya.bhargava and ends at miltiades_official. The specified path is outlined below. The relatively long path of 6 in our graph can be attributed to the fact that dis.copenhagen and yota.katsikouli are not in the strongly connected component.

['divya.bhargava_', 'dis.copenhagen', 'sally.bornhorst', 'zach.annuik', 'swang330', 'yota.katsikouli', 'miltiades_official']

Reciprocity

Reciprocity measures the tendency for edges to be mutual (if A follows B, B also follows A). The overall Graph Reciprocity: 0.4934. This means approximately 49.34% of the existing connections are mutual pairs.

The reciprocity for specific nodes (calculated as the fraction of a node's neighbors involved in a mutual connection with that node) varies, as shown in Table 2.

Note the very low reciprocity for dis.copenhagen, indicating few of the accounts it follows also follow it back, and few of its followers are followed back by the account.

Account	Reciprocity
swang330	0.7385
alex.kalis	0.6460
alexandrapurdy_	0.7316
kabir_aho	0.4171
kirindanek	0.6415
liam_hochman	0.5796
noahpurow	0.5418
seoeunki.m	0.6010
will.deley	0.6172
zach.annuik	0.5571
yota.katsikouli	0.6084
dis.copenhagen	0.0094
elybrayboy	0.5942

Table 2: Reciprocity for individual class members and related accounts.

Subgraph for Simulation (RQ1)

The network we will be using to answer the question - which nodes would be the most effective for a sports team to sponsor if they wanted to grow their brand within our classes DIS students network? - will be a subgraph of our originally constructed network including only nodes which contain at least an out-degree of two. In other words, in order for us to consider a node as a part of our class's DIS students network, they should follow at least two people/accounts from our class. Our network therefore shrunk from 21951 nodes to just 236.

Methods and results

Research question 1: Sponsor simulation

We decided to answer the question - which nodes would be the most effective for a sports team to sponsor if they wanted to grow their brand within our class's DIS students network? - by assuming one of the students/profesor became a sponsored instagram influencer for the sports team, simulating 365 days of our network evolving, and recording the number of nodes/accounts that we can say hold a positive impression of the brand.

Each node in our network is assigned a characteristic titled ‘impression’. The impression refers to the node's relationship to the product in regards to whether the sports team is looked upon favourably and how often the node shares online posts about the sports team to their followers . There are four types of impressions in our model. 1. ‘Sponsor’ - this impression means the node looks favourably upon the team and is required to share three posts about the team a day to their followers. 2. ‘Positive Share’ - this impression means the node looks favourably upon the team, and is engaged to the point that they share one post a week about the team. 3. ‘Positive No Share’ - this impression means the node looks favorably upon the team but does not share posts. 4. ‘None’ - this impression means the node has no favourable view of the team and does not share any of the team's posts. Initially our network contains one chosen node from our Complex Networks class with the ‘Sponsor’ impression, with every other node in the network possessing the ‘None’ impression. The network then goes through our iterable model 365 times, simulating one year. Our model runs the network through both an SI model as well as an Information Cascade model during each iteration. We use a combination of these established models for our simulation due to the multiple characteristics of social media engagement and marketing.

The SI model is used to simulate the probabilistic nature of interactions with shared social media posts. When examining the limited literature regarding how likely someone is to interact and reshare someone else's social media post we found results from small scale experiments carried out in social media blogs Rival IQs “How to Measure and Boost Your Instagram Stories Engagement.” and Crowdfires “How to Get More Retweets.” and when leaning on the positive side of estimations for baseline re-share rates and optimal format multipliers (0.01 % x 2 = 0.02%) we find that the likelihood of followers interacting and resharing an individual post is 1/5000 or 0.02 %. We make the favourable assumption that the very unlikely outcome of interacting and sharing the social media post equates to becoming a fan of the team, which means holding a favourable opinion of the team, and exhibiting fan behaviour around team matches, and since football is the world's most popular sport and on average there is one match a week per team in European football (54 matches), we will assume that each fan shares one post about the team a week. These converted fans hold our ‘Positive Share’ impression. The SI model aspect of our simulation, during every iteration, runs a probability check to convert a node to a

‘Positive Share’ impression for every node that holds a ‘None’ or ‘Positive No Share’ impression that is also a neighbor of a node with a ‘Sponsor’ or ‘Positive Share’ impression where the edge goes to the ‘Sponsor’ or ‘Positive Share’ node and is from the ‘None’ or ‘Positive No Share’ nodes. This means we are running the calculation for instagram followers that view the post from a node that they follow. The final numbers used in the model calculation are different for followers of ‘Sponsors’ and ‘Positive Share’ nodes as the probabilities are adjusted to the frequency of post (3 a day and once a week respectively) with respect to the daily cycle model of our simulation. The SI model iteration is run through at the start of the iteration before the updated Network (after probability checks) is passed to the Information Cascade aspect of our simulation.

The Information Cascade model of our simulation simulates the longer-term and more consistent aspects of social media marketing where positive brand association is derived from the shared opinions of those whom one follows over a prolonged period of time. When examining how one is influenced by the recommendations of a trusted circle (which we are assuming our DIS network is), when using the Diffusion of Innovations theory from Everett M. Rogers, *Diffusion of Innovations* (5th ed., Free Press, 2003), we find that on average low-involvement products, medium-involvement products, and high-involvement products require recommendation from 15%, 50%, 75%, of the trusted circle respectively to be adopted/engaged with. We ran simulations with these three different thresholds, where “viewing the team favourably” was defined as being willing to purchase/engage with a team branded product that matches the commitment (be it financial or otherwise) corresponding to the chosen favorability threshold. We say that a node is above that threshold by comparing the threshold to the % of neighbours to our chosen node that have either a ‘Sponsor’ or ‘Positive Share’ impression, and are connected by an edge starting from the chosen node and going to the neighbour. This means we are comparing the % of nodes that our chosen node follows, that post about the sports team, to our chosen threshold value. We run this check for every node with the ‘None’ impression and update them to the ‘Positive No Share’ impression if the threshold value is reached or exceeded. This updated graph is then returned before it then begins its next iteration which begins with the SI model aspect of our simulation before entering the Information Cascade aspect of our simulation again until the 365 iterations are complete.

Our model returns the total nodes possessing each of the four impressions. These statistics, and specifically the total number of positive impressions (Sponsor + Positive Share + Positive No Share), are used to judge the effectiveness of each simulation attempt's specific sponsor from our Complex Networks Class.

Results

Below you will see the graphs depicting the effectiveness of each sponsored influencers year in regards to how many people in the Complex Networks DIS student Network hold a positive impression of the sports team brand, for each of the three unique thresholds used, along with a visualization of the network from the simulation with the most effective sponsor for each threshold. The colors on the visualization correspond to the following impression: Blue = Sponsor. Dark Green = Positive Share. Light Green = Positive No Share. Red = None.

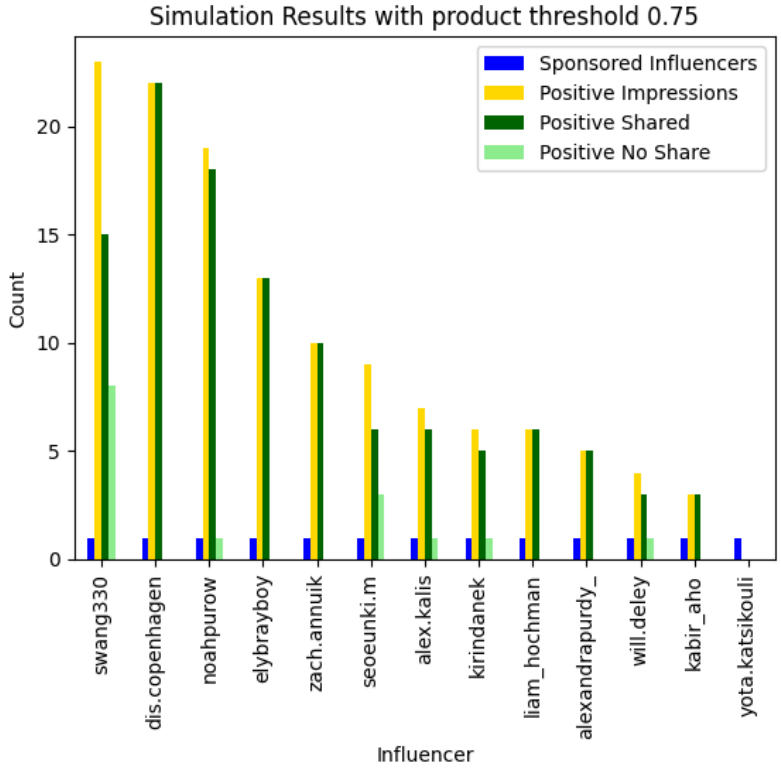

Figure 3

Figure 4

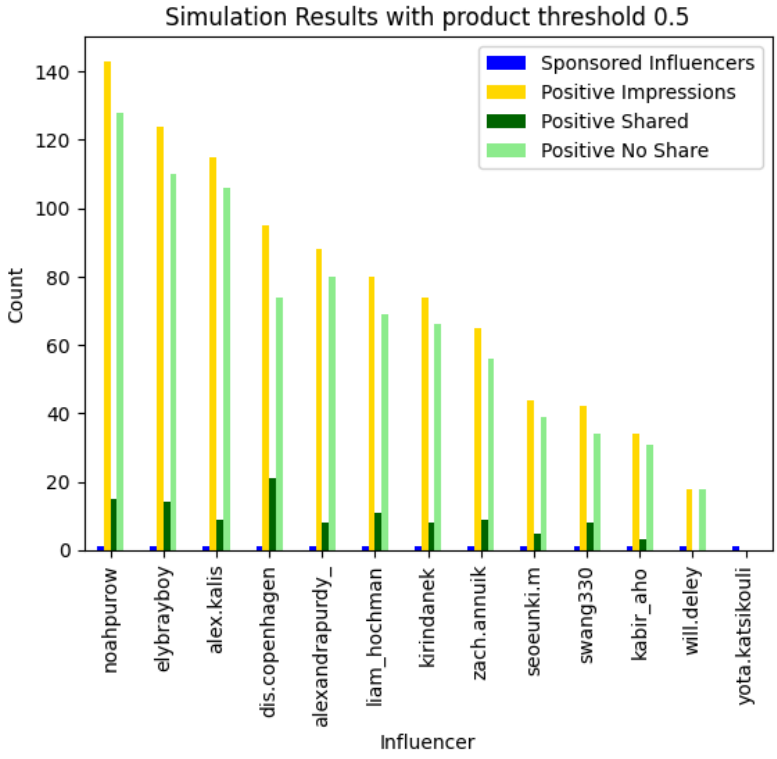

Figure 5

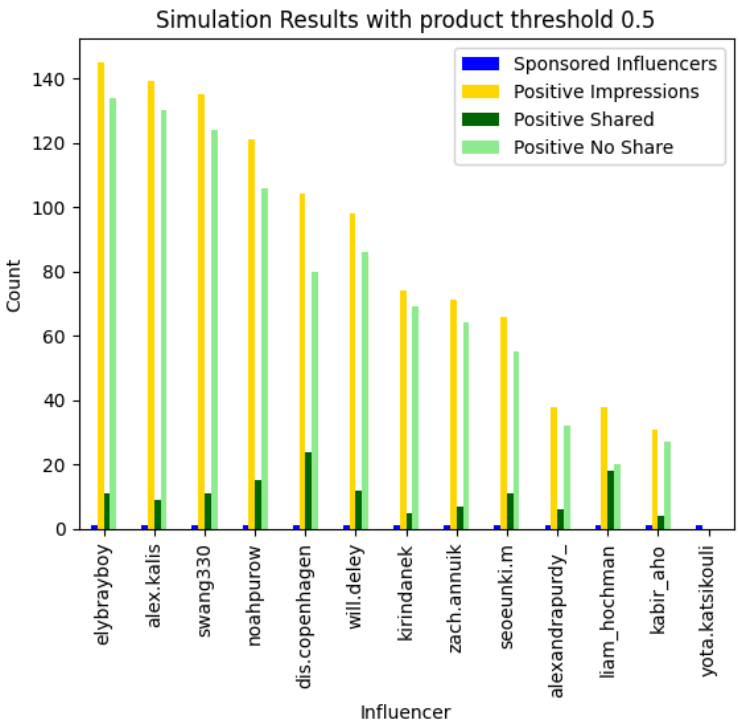

Figure 6

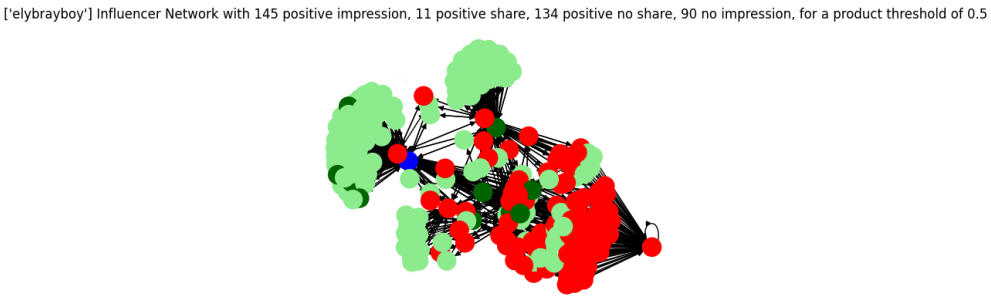

Figure 7

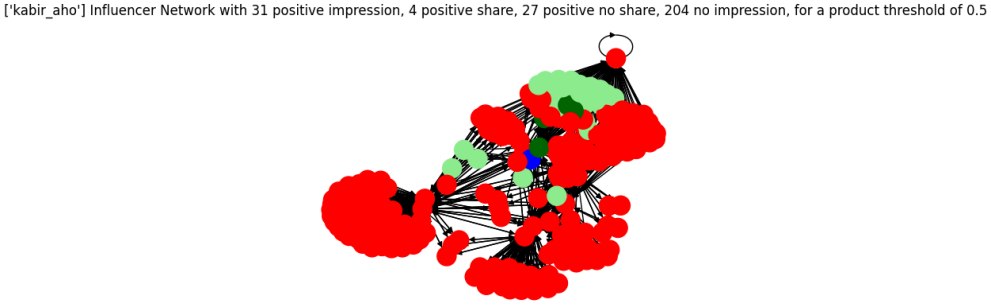

Figure 8

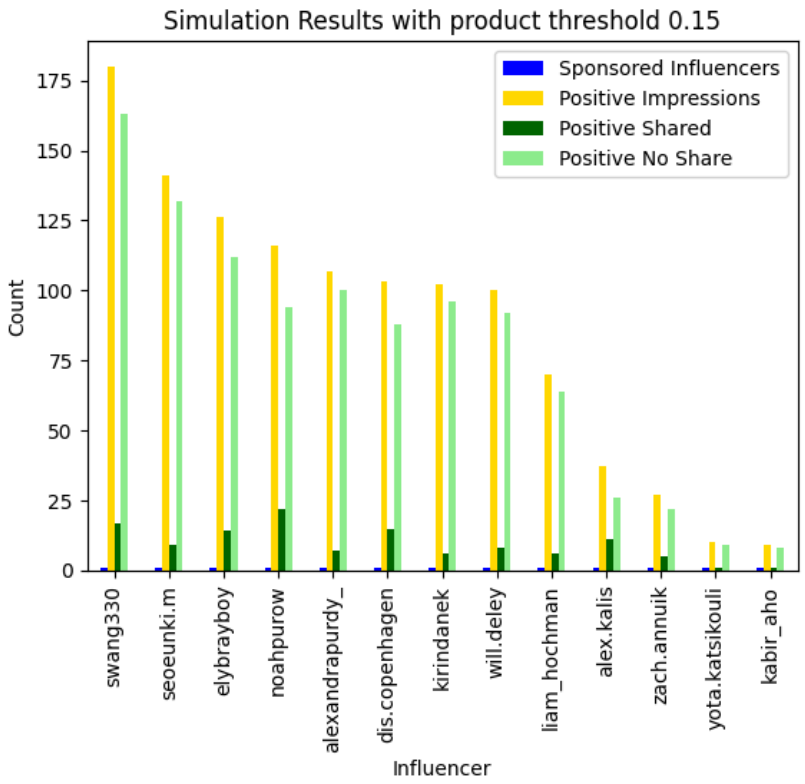

Figure 9

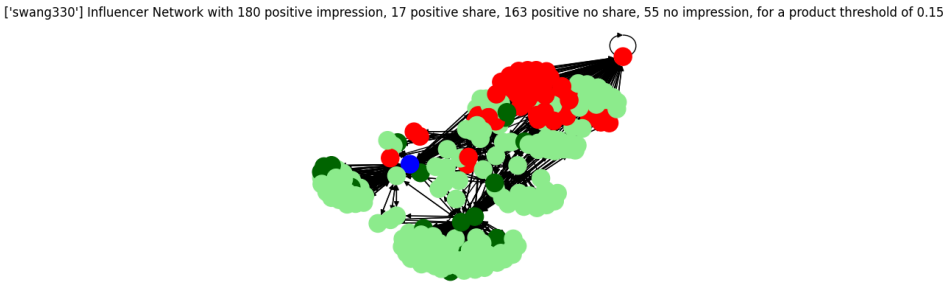

Figure 10

Our results for each threshold largely met expectations in regards to total counts. There is a clear difference between influence for high threshold products (high of 23 from swang330 - figure 3), compared to the high thresholds for the medium and low threshold products (145 from elybrayboy - figure 6 - and 180 from swang330 respectively - figure 9). What was interesting was how little the difference in distribution between the low threshold and medium threshold peaks were. This could be explained by most nodes having an in degree of either two or 3,

meaning just a one node difference lay between the thresholds for both nodes. When running these simulations for different thresholds, the incredibly high variance in regards to which accounts would be the most influential was noticeable. This is no doubt due to the probabilistic nature of social media posts being shared by others, and shows the high variance of the influencer marketing model. We however find groups of accounts that consistently perform relatively poorly, such as kabir_aho, and yota.katsikouli, explained by their relatively low out-degree in our filtered network. The rest of the accounts have the capacity to spread high positive influences due to their interconnectedness to the giant component, but still even these nodes retain the capacity to not spread many positive impressions (as seen by swang330s performance in the first 0.5 threshold graph - figure 5, - compared to their strong performance in the other three simulations shown). We found the distributions of different simulations of the same threshold to be pretty consistent (as seen with the example of the very close distributions for two different 0.5 threshold simulations - figures 5 and 6), even if individuals were not, suggesting sponsors should invest into a market through investing in multiple high degree nodes rather than an individual influencer for a more consistent return. This information is really insightful not just for the context of the network derived from our class, but for understanding the social media influencer market as a whole.

Research Question 2: Connecting Accounts

Approach

Our goal is to identify connecting Instagram accounts that are not in the class. We consider ‘connecting’ accounts to be accounts that have (possibly non-direct) relationships with many class members when class member to class member relationships are taken away.

Our first thoughts on how to approach identifying connecting accounts focused on identifying the non-class member accounts with the highest in-degree. Since each directed edge in our network u -> v with v belonging to the set of non-class accounts represents a class member following a non-class account, non-class members with high in-degrees are followed by multiple class members, and as such, show a relationship between class members. However, this approach yielded two mutually exclusive problems. The first problem is that celebrities and group

accounts (the accounts with the highest in-degrees in the network graph are DIS Abroad, Kanye West, Livingston Football Club, and LeBron James) were higher than any people that class members might know personally. The second problem (a more serious issue, in our eyes) is that while this approach identifies the accounts that are followed by the highest number of class members, it has the potential to totally exclude class members who don't know the popular accounts. If a class member doesn't follow a high-in-degree account, there is nothing in our strategy that encourages the "connecting" accounts to be connected in any way with them. The first issue is easily solvable by considering out-degree instead of in-degree; however the second is not.

Hence, we set out to find a middle ground that balances the criteria of (1) the ability to reach a large number of class members with (2) the ability to be reached from a large number of class members. With these refined criteria in mind, our eyes turned to HITS hubs score and betweenness centrality. While the former improves on our first attempt by taking into account more complex relationships, we still run into the same second problem as our previous attempt where highly-ranked accounts can be highly disconnected from some class members. The latter, meanwhile, is not helpful since nearly all class members are connected via direct edges.

However, these early approaches did identify a path to follow to identify central accounts. To summarize our process, we first partition our follower graph G=(V,E) into two sets: L: Nodes representing class members, and R: Nodes representing non-class members. From there, we remove all edges connecting class members, and introduce synthetic edges E' connecting some non-class members using a method outlined by Zhao et al. (2007). Finally, we augment the graph with another layer L'=L.copy(): A copy of L, and remove all edges R -> L and L' -> R. This yields a graph T=(L U R U L', E U E') such that (L U R)=V and all paths from L -> L' are of the form L -> R (-> R->...) -> L'.

Given this three-partition graph, we are able to compute betweenness centrality on the subset of paths that begin in L and end in L', essentially calculating which nodes are most central to members when paths are not allowed to traverse class members.

While this is not a perfect solution to our two criteria for "connecting" accounts, we believe it is a promising approach that merits further exploration. We demonstrate its implementation and results, and show how it differs from the basic calculations of in-degree and HITS hub scores.

Methods

Step 1: Create bipartite G: We are given our network graph G=(V,E). We partition V into two sets: L: (v in V | v is a class member) and R: (v in V | v is not a class member). We delete all edges ((u, v) | u in L AND v in L), and are left with a bipartite graph where the only remaining edges are between class members and non-class members.

Step 2: Augment E with synthetic edges: We now wish to introduce directed edges connecting nodes in R to approximate how non-class member following relationships might look if we were able to observe them. This is a difficult task, but we look to the existing literature on bipartite network projection and modify a technique proposed in *Bipartite Network Projection and Personal Recommendation* by Tao Zhou et al., 2007. This technique uses common neighbors in L to create a projection of the partition R of a bipartite graph B=(L U R, E). A bipartite graph projection is a set of weighted edges between nodes in the same bipartite partition where the weights aim to represent the structure of the original bipartite graph. A simple approach is displayed in (Figure [Reference to zhou1]), where edge weightings (x1, x2) in the bipartite graph G=(X U Y, E) are simply the number of shared neighbors of nodes x1 and x2.

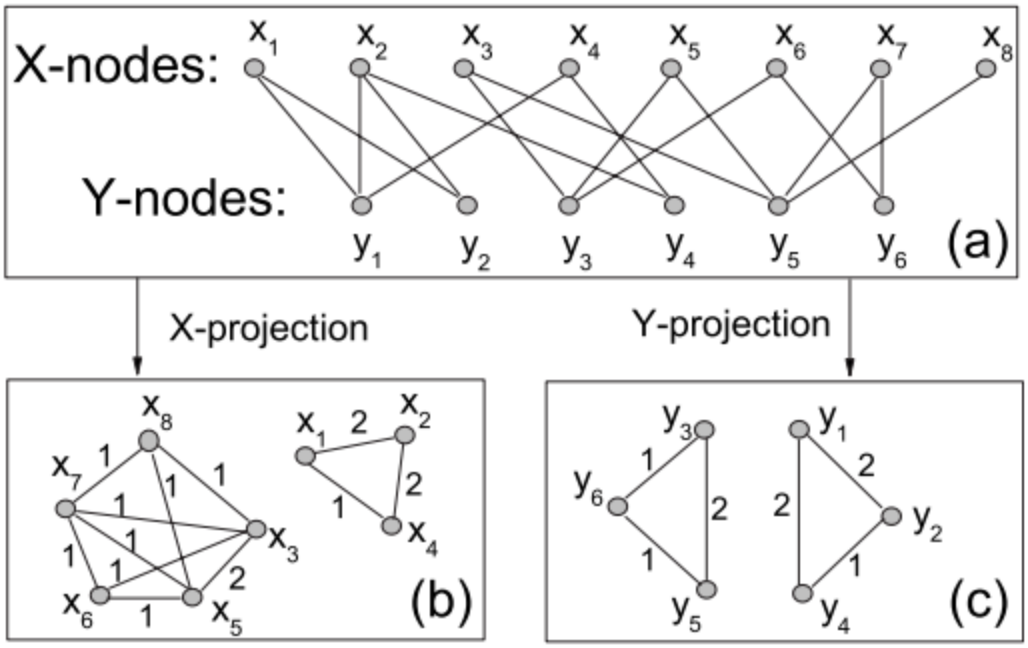

Figure 1 from Zhou et al. (2007): Illustration of a bipartite network (a) as well as its X projection (b) and its Y projection (c). Edge weights in the projected graphs are the number of shared neighbors.]

We take a more nuanced approach, as outlined in Zhou et al. (2007). Specifically, we conduct a resource-allocation procedure to assign weights to a fully connected graph G_R=(R, W) where weights w_{i,j} in W represent how much "resource" can flow from a node u in L to be shared by nodes i,j in R. The procedure is as follows:

Algorithm Pseudocode:

```
For all {i,j} in R:
  If i==j continue
  w_{i,j}=0 // add weighted directed edge i->j with weight 0
End for


For each l in L:
  N[l]=(set of nodes in R that l has a directed edge to)
  For each ordered pair of nodes (u,v) in N[l]:
    w{u,v}+=1/(k_out(l) * k_in(v))
```

End for
End for

Once we have calculated all edge weights, we select only edges with weights > τ (tau) to keep as directed edges in our final graph, where τ (weight threshold) is a hyperparameter. We take inspiration from the K-means approaches demonstrated in class and recommend an "elbow method" implementation by KneeLocator to select τ when accounting for outliers. Unfortunately, due to memory constraints we were unable to run KneeLocator on our full set of weights, and used a truncated set of the top 11,000,000 weights to locate an "elbow", yielding a somewhat arbitrary tau value of τ = 0.001441.

We add the pruned set of directed edges W to the originally-bipartite graph G.

Step 3: Create third partition L': Next, we add a third partition of the nodes of G, L'. We initialize L' as an exact copy of all nodes in L, where each l' in L' has edges to and from the same nodes in R as its corresponding node in L.

We then discard any edges R -> L and L' -> R, so that the only remaining paths from L -> L' follow the form L -> R -> (R -> ...) -> L'.

Step 4: Compute betweenness centrality: Finally, we simply compute the betweenness centrality of the finished graph on the subset of paths that begin in L and end in L'.

Results

Our final three-partition graph contained 21,964 total nodes. Of these nodes, 13 were in the set L; 21938 in the set R; and 13 in the set L'. Our graph contained 119,835 total edges, of which 90,006 were R->R (synthetic); 15,781 were L->R; and 14,048 were R->L'.

The top ten R-nodes (non-class members) by L->L' betweenness were:

Node	Betweenness	InDeg	OutDeg
disabroad	1.2439e-08	7	182

```
  anderss.k              5.2203e-09      2      4
  itzkim_y               4.4203e-09      2      3
  will.b.lambert         4.3782e-09      2      3
  sally.bornhorst        4.2365e-09    239    240
  lylahaikoye            4.1543e-09      2      2
  _anders_hansen         4.1529e-09      3      2
  mindy.kimm             4.1524e-09      2      2
  shaoyxx                4.1521e-09      2      2
  3liana_mn              3.5787e-09      3      4
```

We attempted to evaluate these metrics by checking the shortest directed paths from class members to some of these connecting accounts, and achieved the following results:

```
 will.deley   -> anderss.k    :  no path
  will.deley   -> sally.bornhorst :  no path
  yota.katsikouli -> anderss.k    :  no path
  yota.katsikouli -> sally.bornhorst :  no path
  noahpurow    -> anderss.k    :  no path
  noahpurow    -> sally.bornhorst :  noahpurow  ->
tuftsuniversity  ->  sally.bornhorst
  kabir_aho    -> anderss.k    :  kabir_aho  ->  anderss.k
  kabir_aho    -> sally.bornhorst :  kabir_aho  ->  fordsabroad
->  sally.bornhorst
  alex.kalis   -> anderss.k    :  no path
  alex.kalis   -> sally.bornhorst :  no path
  seoeunki.m   -> anderss.k    :  no path
  seoeunki.m   -> sally.bornhorst :  no path
  elybrayboy   -> anderss.k    :  no path
  elybrayboy   -> sally.bornhorst :  no path
  liam_hochman -> anderss.k    :  no path
  liam_hochman -> sally.bornhorst :  no path
  swang330     -> anderss.k    :  no path
  swang330     -> sally.bornhorst :  no path
  alexandrapurdy_ -> anderss.k    :  no path
  alexandrapurdy_ -> sally.bornhorst :  no path
  dis.copenhagen -> anderss.k    :  no path
  dis.copenhagen -> sally.bornhorst :  dis.copenhagen  ->
sally.bornhorst
  zach.annuik  -> anderss.k    :  no path
  zach.annuik  -> sally.bornhorst :  no path
  kirindanek   -> anderss.k    :  kirindanek  ->  anderss.k
```

```
  kirindanek   -> sally.bornhorst :  kirindanek  ->  uwmadison
->  sally.bornhorst
```

Also, interestingly, no class members in this construction (aside from Yota herself) had paths to Yota.

Evaluation: Our graph still has few paths from most class members to non-class members. We wonder whether this is because we really aren't that connected, or if it's because we failed to add enough, or useful enough, R-R synthetic edges. We also note that our algorithm has given a few nodes in R very high out-degrees, and the rest virtually no additions. Another problem with our synthetic edges R-R is that group or celebrity accounts such as university accounts ("tuftsuniversity", "fordsabroad") are set to follow individuals, which is unlikely to occur in real life. In the future, we would like to refine these synthetic edges by better considering follower/following relationships in the bipartite network projection– for example, creating weight contributions between (u,v) in R that are proportional to how much resource can flow from u to v through an intermediate node in L instead of how much resource can flow from a node in L to u and v.

Finally, we compare all methods mentioned in this section in the following table:

In-Degree	**In-Degree (no group/celebs)**	**HITS Hubs in B**	**Betweenness Centrality in T**
DIS Abroad	baybater	3liana_mn	DIS Abroad
Kanye West	3liana_mn	will.b.lambert	anderss.k
Livingston FC	guslafave	anderss.k	itzkim_y

Our initial approach of in-degree calculations presented the problems of celebrities and non-individual accounts alongside the problem of possible high disconnections. Removing celebrities and non-individual accounts explicitly solved the first problem, but not the second. HITS hubs calculated in the bipartite graph B (no L-L connections) solved the first issue, and

improved upon the second. Betweenness centrality computed on our three-partition graph T proved to be the most effective at balancing these two issues, but was still problematic due to potentially spurious R-R synthetic edges. Overall, we identify two accounts that consistently arise: ‘3liana_mn’ and ‘anderss.k’.

Research Question 3: Analyzing Asymmetric Cycles of Length 3

Motivation

Beyond direct connections and overall network centrality, the local structure surrounding nodes offers insights into their roles within a social network. One fundamental local structure is the simple cycle of length three. In directed networks like Instagram follower graphs, these cycles can be either fully reciprocal (symmetric: A follows B, B follows A, A follows C, C follows A, B follows C, C follows B) or contain non-reciprocal relationships (asymmetric: e.g., A follows B, B follows C, C follows A, but B does not follow A).

It’s often considered an awkward social situation when someone you follow doesn’t follow you back. Exploring this phenomenon at the local level can reveal more nuanced social dynamics. Symmetric cycles of length 3 represent tightly-knit groups where everyone mutually follows each other, suggesting strong cohesion. Asymmetric cycles, however, indicate directed flow or influence hierarchies within the local group. For example, A -> B -> C -> A might represent a situation where information or influence potentially flows in a cycle, but the relationships aren't all mutual. Understanding which nodes participate frequently in these cycles, and the balance between symmetric and asymmetric cycles they are involved in, can help characterize their potential roles (e.g., core member of a clique vs. bridge).

Therefore, our third research question focuses on analyzing these 3-cycles within the core of our network (the largest strongly connected component, LSCC): Which nodes within the LSCC participate most frequently in simple cycles of length 3? How does the proportion of asymmetric versus symmetric cycles a node participates in characterize its local network role and potential role in reciprocal or directed information flow in the network as a whole?

Methods

To investigate the role of nodes within the smallest cyclic structures, we focused on simple cycles of length 3 within the largest strongly connected component (LSCC) of the directed graph, which contains 7278 nodes. The analysis involved the following steps:

Cycle Enumeration: We identified all unique simple cycles of length 3. This was done by iterating through each node u in the LSCC. For each neighbor v of u (i.e., an edge u -> v exists), we checked if any neighbor w of v (i.e., an edge v -> w exists) also had an edge pointing back to the original node u (i.e., w -> u). If such a path u -> v -> w -> u existed, and w was not equal to u, it formed a 3-cycle. To ensure uniqueness, cycles were stored using a canonical representation (e.g., sorting the nodes lexicographically).

Asymmetry Classification: Each identified 3-cycle (u, v, w) was classified as either symmetric or asymmetric. A cycle is symmetric if all adjacent pairs of nodes within the cycle have reciprocal edges within the LSCC (i.e., u <-> v, v <-> w, and w <-> u all exist). If any one of these pairs lacks a reciprocal edge (e.g., u -> v exists but v -> u does not), the cycle is classified as asymmetric.

Node Participation Count: We counted the total number of unique 3-cycles (both symmetric and asymmetric) each node participated in.

Asymmetry Proportion Calculation: For each node, we calculated the proportion of the 3-cycles it participated in that were asymmetric. This proportion is calculated as (Number of Asymmetric 3-Cycles Node Participates In) / (Total Number of 3-Cycles Node Participates In). This metric ranges from 0 (all cycles are symmetric) to 1 (all cycles are asymmetric).

Results

Our analysis of 3-cycles within the Largest Strongly Connected Component (LSCC, 7278 nodes) yielded the following results:

Overall Cycle Counts: We found a total of 380 unique simple cycles of length 3 within the LSCC. Of these, 120 were classified as asymmetric.

Overall Asymmetry Proportion: The overall proportion of asymmetric 3-cycles in the LSCC is 120 / 380 ≈ 0.3158. This indicates that roughly 31.6% of the 3-cycle structures within the network's core involve at least one non-reciprocal follower relationship.

Node Participation: Nodes varied significantly in the total number of 3-cycles they participated in. (The full list of participation counts per node is omitted for brevity but is available in the analysis notebook). Identifying the top participating nodes highlights individuals who are most central to these local cyclic processes within the LSCC.

Asymmetry Proportion per Node: The asymmetry proportion calculated for each node reveals characteristics of its local structural role. Table 3 shows these proportions for the class members and related accounts within the LSCC. The in-degree and out-degree listed are also calculated within the LSCC subgraph.

Account	Asymmetry Prop.	LSCC In-Deg	LSCC Out-Deg
elybrayboy	0.4214	708	723
noahpurow	0.3709	651	684
liam_hochman	0.3146	724	739
alex.kalis	0.2787	945	951
swang330	0.2376	662	674
kirindanek	0.2340	758	768
zach.annuik	0.2075	475	471
kabir_aho	0.2000	662	670
will.deley	0.1967	414	413
seoeunki.m	0.1613	764	771
alexandrapurdy_	0.1429	579	588
yota.katsikouli	0.0000	101	101
dis.copenhagen	0.0000	156	24

Table 3: Asymmetry Proportion in 3-Cycles for Class Members (within LSCC).

Interpretation

The analysis of 3-cycles provides a lens into the local structural dynamics within the network's core (LSCC).

The overall asymmetry proportion of ~31.6% suggests that while mutual relationships are prevalent in forming symmetric cycles, a significant portion of these local cycles incorporate directedness, hinting at hierarchical or non-uniform influence patterns even within this strongly connected component.

Examining the asymmetry proportion for individual nodes allows for their characterization:

Nodes with a low proportion (closer to 0) primarily exist within fully reciprocal symmetric cycles. These nodes likely belong to tightly-knit, mutually connected local groups where relationships are predominantly mutual. yota.katsikouli and dis.copenhagen have a proportion of 0, meaning all 3-cycles they are part of within the LSCC are fully symmetric. This aligns with their somewhat distinct roles – yota.katsikouli being an instructor and dis.copenhagen being an organizational account – suggesting their cyclic involvement is within groups where all members mutually follow each other (within the context of the LSCC). Other class members like alexandrapurdy_, seoeunki.m, and will.deley also show relatively low proportions, suggesting stronger embedding in mutual cycles.

Nodes with a high proportion (closer to 1) frequently participate in cycles where influence or connection is not fully mutual (asymmetric cycles). These nodes might act as bridges connecting different local subgroups or be part of directed information/influence flow paths within the LSCC. elybrayboy (0.421) and noahpurow (0.371) exhibit the highest asymmetry proportions among the class members, suggesting their local cyclic connections more often involve non-reciprocal links. liam_hochman (0.315) is close to the network average.

This analysis complements other centrality measures by focusing specifically on the nature of the smallest cyclic structures a node is embedded in. It helps differentiate nodes that might have similar overall connectivity but play different roles in local information flow and group cohesion based on the reciprocity within their immediate cyclic neighborhood.

Conclusion

We have collected information about Instagram follower-following relationships for 11 students in our Complex Networks class; our 1 instructor; and the DIS Copenhagen account. We obtained our data by scraping Instagram, and then created our social network using the NetworkX library. We used this network to investigate structural aspects of our real-world social relationships, and explored influence spread, connectedness of class members, and the interesting phenomenon of network 3-cycles.

Our marketing campaign simulation applied an SI and Information Cascade model to determine which class members could be the most effective sponsored influencers. Our connecting accounts analysis proposed a novel application of a bipartite graph projection method

designed by Zhou et al. (2007) for use in reconstructing missing non-class relationships, and used this to compute betweenness centrality on an augmented graph. Finally, our analysis of asymmetric 3-cycles within the LSCC of the network revealed the potential existence of different modes of local information flow in cyclic neighborhoods

Overall, our social network demonstrates powerful analytical properties, but has significant limitations due to not measuring follower-following relationships between non-class members. While Instagram social networks are a promising method for identifying and analyzing social relationships, complete data collection is difficult. Future work should focus on enhancing synthetic graph construction or expanding data collection.

Works Cited

Crowdfire Inc. 2021. “How to Get More Retweets - the Crowdfire Blog.” The Crowdfire Blog. April 27, 2021. https://read.crowdfireapp.com/2021/04/27/how-to-get-more-retweets/.

Gualtieri, Jackie. 2023. “How to Measure and Boost Your Instagram Stories Engagement.” Rival IQ. May 24, 2023. https://www.rivaliq.com/blog/instagram-stories-engagement/.

Rogers, Everett. 2003. “DIFFUSION of INNOVATIONS Third Edition Library of Congress Cataloging in Publication Data the AMERICAN CENTER LIBRARY Contents.” https://teddykw2.wordpress.com/wp-content/uploads/2012/07/everett-m-rogers-diffusion-of-innovations.pdf.

Zhou, Tao, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. 2007. “Bipartite Network Projection and Personal Recommendation.” *Physical Review E* 76 (4). https://doi.org/10.1103/physreve.76.046115.