

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Development of a System to Collect Social Network Data from College Students for Future Studies in Health Behavior and Academic Performance /

Permalink

<https://escholarship.org/uc/item/2dv10709>

Author

Lah, Mike Myoungwhan

Publication Date

2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

**Development of a System to Collect Social Network Data from College
Students for Future Studies in Health Behavior and Academic
Performance**

A thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Mike Myoungwhan Lah

Committee in charge:

Professor Lucila Ohno-Machado, Co-Chair
Professor Charles Elkan, Co-Chair
Professor Kamalika Chaudhuri

2013

Copyright

Mike Myoungwhan Lah, 2013

All rights reserved.

The thesis of Mike Myoungwhan Lah is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Co-Chair

University of California, San Diego

2013

TABLE OF CONTENTS

Signature Page		iii
Table of Contents		iv
List of Figures		vi
Acknowledgements		vii
Abstract of the Thesis		viii
Chapter 1	Background	1
	1.1 Motivation	1
	1.2 Contagion in Social Networks	2
	1.3 Isolating Contagion Effects	2
	1.4 Inferring Tie Strength from Facebook Interactions	3
Chapter 2	Methods	4
	2.1 Legal Considerations	5
	2.1.1 IRB	5
	2.1.2 HIPAA	5
	2.2 TritonSchedule	5
	2.2.1 Application Architecture	6
	2.2.2 Signup Flow	6
	2.2.3 Features	10
	2.2.4 Benefits of our Study	11
	2.3 Gathering Facebook Data	11
	2.4 Gathering Academic Data	12
	2.5 Technical Challenges	12
	2.5.1 Facebook API Limit	12
	2.5.2 Data Size	13
	2.5.3 Facebook API Changes	13
Chapter 3	Results	14
	3.1 General Statistics	14
	3.2 Inferring Tie Strength	15
Chapter 4	Discussion and Future Work	18
	4.1 Future Work	19
	4.1.1 Combination with Network Security	19
	4.1.2 Combination with Health Behaviors	19
	4.1.3 Natural Language Processing	19

4.2	Lessons Learned	20
4.2.1	IRB Review Process	20
4.2.2	Accessing Student Data	20
4.2.3	Development in a Secure Environment	21
4.2.4	Advertisement and Growing the Study Population	21
Appendix A	Informed Consent Form	23
Bibliography	26

LIST OF FIGURES

Figure 2.1:	The TritonSchedule login page. The user sees pictures of their Facebook friends who are already using TritonSchedule.	7
Figure 2.2:	The consent page. This is shown before the Facebook Authorization dialog, so the user is required either to accept the terms of the study, or choose not to participate. See Appendix A for the complete consent form.	8
Figure 2.3:	The Facebook Authorization dialog. This is a mechanism provided by Facebook for users to either authorize applications access to certain parts of their Facebook profile, or choose not to use the application. In the dialog, TritonSchedule asks for access to a user’s News Feed.	8
Figure 2.4:	The UCSD Single Sign-On page. Users must verify their UCSD identity before using TritonSchedule.	9
Figure 2.5:	The main TritonSchedule page. Students can see friends who are in the same courses, given that the friends have also signed up for TritonSchedule.	10
Figure 3.1:	The largest connected component of the social network of around 180 TritonSchedule users. The author is highlighted in yellow. . .	15

ACKNOWLEDGEMENTS

I would like to thank my adviser Professor Lucila Ohno-Machado for her constant guidance, insight, encouragement, and support. I am also grateful to Professor James Fowler, and Dr. Karen Calfas, who provided the concept for this project, and the experience with the administration necessary for our success.

Many thanks to the staff at the UCSD Registrar's Office, ACT, CTRI, and SDSC. They put a lot of work into setting up the infrastructure for this project, and none of it would have been possible without them.

Thank you to Professor Charles Elkan and Professor Kamalika Chaudhuri for generously agreeing to read this thesis, and for their invaluable feedback.

For their generous support of this project, my sincerest gratitude to iDASH and the NIH Grant U54HL108460 to the University of California, San Diego.

ABSTRACT OF THE THESIS

Development of a System to Collect Social Network Data from College Students for Future Studies in Health Behavior and Academic Performance

by

Mike Myoungwhan Lah

Master of Science in Computer Science

University of California, San Diego, 2013

Professor Lucila Ohno-Machado, Co-Chair
Professor Charles Elkan, Co-Chair

Researchers study social networks to understand how individuals with similar behavior form clusters, and what causes them to do so. Universities are interested in learning more about influential factors of student behavior, including the impact that their social networks have on these behaviors. We have done foundational work to collect a dataset about UCSD student social network data gathered from Facebook and academic data from the UCSD Registrar. Once complete, the social network portion of this dataset will also be combined with datasets on health behaviors, potentially to build predictive models for depression, substance abuse,

and other important conditions.

Chapter 1

Background

1.1 Motivation

Humans are social creatures. We form like-minded groups, and recommend books and restaurants to our friends. We can easily observe clustering in social networks, but it is often difficult to determine whether the clustering is a result of homophily (“love of the same”) or contagion (individuals influencing each other) [2, 6]. Research suggests that some level of contagion occurs in social networks, and that the effect extends up to three degrees (friends of friends of friends) [6]. The social network, especially the closest members, can have a strong influence on the health and happiness of an individual [6].

Universities are very concerned about the health and well being of their students, especially related to factors such as substance abuse, anxiety, and depression [1]. We believe that by examining these factors in the context of the social network, we will develop a better understanding of how students influence each other’s behavior and mental state. We hope that this understanding will allow university administrators develop innovative ways to improve student health and academic performance. Researchers are exploring intervention models based on online social activity, which leverages support and influence in the social network, and could be an effective model for reaching students [3, 10, 11].

We are building a data set that will allow us to explore correlations between students’ health and academic performances. We recruit students into our study

through a Facebook application, which is also how participants provide their consent. We gather social network data through the Facebook API, and get academic records through the Registrars Office. This document is a summary of the planned analysis, progress on compiling the data set, and lessons we have learned so far.

1.2 Contagion in Social Networks

Fowler and Christakis have shown that certain behaviors (obesity, smoking, happiness) spread throughout the social network [4, 5, 6] by studying longitudinal data gathered as part of the Framingham Heart Study. They found that if a person's friend became obese, started smoking, or became happy, they became more likely to develop that behavior as well. This effect was significant up to three degrees in the social network (friends of friends of friends) [6].

Clustering is easily observed in a social network, but it is difficult to determine the cause of that clustering. Generally, there are three reasons for clustering in a social network: homophily, confounding, and contagion. Homophily is the tendency for similar individuals to group together; birds of a feather flock together. Confounding is a result of the shared experience of an external event; if there is a natural disaster in an area, the residents in that area will become unhappy. Contagion is the influence that individuals have on each other; a causal relationship between an individual and their friends [2, 6]. We are interested in both observing clustering in the student social network, and trying to isolate contagion effects.

1.3 Isolating Contagion Effects

To isolate contagion effects in the social network, Fowler and Christakis ran multiple regression models of individual behavior, using current and past behavior of the individual and their friends as features [6]. They considered several factors in trying to isolate causal effects, including the following:

Alter's happiness in the previous exam helps to control for homophily. We evaluated the possibility of omitted variables or contemporaneous events or exposures in explaining the associations by examining how

the type or direction of the social relationship between ego and alter affects the association between them. If unobserved factors drive the association between ego and alter happiness, then directionality of friendship should not be relevant. We also examined the possible role of exposure to neighbourhood factors by examining maps. [6]

These statistical techniques rely on long-term observation of both the individual features (happiness, etc.), as well as the existence and direction of edges in the graph (who values whose friendship more or less). We made sure to collect enough long-term data about users to make this kind of longitudinal analysis possible.

1.4 Inferring Tie Strength from Facebook Interactions

Users of Facebook can interact with their friends in multiple ways, including posting public messages (wall posts), commenting on content, “liking” content (commonly seen as a thumbs up icon), tagging friends in photos, and sending private messages. The Facebook News Feed shows a summary of posts by the user’s friends, and the associated public interactions. Jones et al. found that a simple logistic regression model using a single feature, the sum of interactions between a user and their friends, was 82% accurate at predicting whether a user considered that friend a “closest friend”. In addition, they found that private communication (private messages) was not necessarily more predictive than public communication (wall posts) [13]. This suggests that we will be able to use public communication available through the Facebook Graph API (namely, the News Feed) to infer directed tie strength between users.

Previous work on tie strength has required users to come into the lab [8], or direct cooperation with Facebook [13] to gather the data necessary to make inferences. We believe our study is novel in that we gather all Facebook data through the public Facebook API. This gives us the potential to recruit more users over time, and the Facebook application and data gathering portion of our work is easily portable to other institutions.

Chapter 2

Methods

Our proposed dataset will consist of a directed graph, where the nodes are UCSD students, and the edges are the Facebook friend connections between students. The edge weights are the number of interactions between students, and this allows evaluation of tie strength between students (which students are closest friends in real life). We gathered a variety of academic and demographic data on students in our study, such as GPA, hometown, current address, and high school academic performance. We plan to investigate which of these factors is most correlated with student interactions.

There are three main components to developing this dataset: 1) developing a Facebook application to recruit users into our study, 2) gathering Facebook data using the Facebook API, and 3) gathering academic data from the UCSD Registrar. This master's thesis primarily reports on item (1) above, including work preparatory for the final research involving social networks and student characteristics, which includes institutional approval of the application and research protocols, development of the application within the legal requirements for handling protected health information, and negotiation with UCSD administrators about how to obtain data from the registrar's office.

2.1 Legal Considerations

Before conducting any research involving human subjects for this study, there were two important requirements that we had to satisfy: (a) obtaining approval from the Human Research Protection Program (HRPP), often referred to as the Institutional Review Board (IRB) [16], and (b) compliance with the Health Insurance Portability and Accountability Act (HIPAA) [14].

2.1.1 IRB

All human subjects research at UCSD is subject to review by the IRB. We had to submit a detailed research plan and obtain IRB approval (after several revisions) before releasing TritonSchedule to students and collecting student data. See Appendix A for the complete informed consent form approved by the IRB. All research personnel were certified to participate in research involving human subjects.

2.1.2 HIPAA

HIPAA is a law that, among other things, regulates the storage and handling of health related PII (Personally Identifiable Information). Over the course of our study, we collected PII, so we had to ensure that our servers were HIPAA-compliant. This included firewalls, access control, and physical security requirements [14]. Some of these security considerations are detailed in Section 2.2.1 Application Architecture.

2.2 TritonSchedule

We had three main reasons for developing a Facebook application: streamlining the signup to participate, getting authorization to access Facebook data, and providing a useful service to students that would keep them in the study and allow us long term access to their Facebook data. We developed a Facebook application

called TritonSchedule that allowed UCSD students to share their course schedule with their friends who are also using TritonSchedule.

2.2.1 Application Architecture

A Facebook application is a website or web application that is registered with Facebook, and displayed on their site through an HTML `iframe` element. TritonSchedule was written using PHP, HTML, and Javascript, and runs on an Apache web server. MySQL is used for data storage.

As data security is one of our top priorities, we use some additional security measures. Additional security for the application server is provided by a proxy server, which lives on the DMZ. The application server does not service user requests directly. All requests come to the proxy server, which is configured as an Apache reverse proxy to the application server. All servers are accessible only through two-factor authentication.

To reduce our risk profile, we also created separate databases for application data and research data. Application data consists of only the data required for TritonSchedule to function. Academic data includes basic individual data (name, PID) and course schedule data. Research data is more sensitive personal data we will use only for research, and is not visible to TritonSchedule users. Research data includes Facebook News Feed data and academic records, such as grades. We created an additional research server that is used for all gathering and parsing of research data, and restricted the application server to only have access to the application database. In the event that the TritonSchedule application is compromised, an attacker would only have access to the less sensitive application data.

2.2.2 Signup Flow

Signup for TritonSchedule, and our study, proceeds in two steps. The first step is obtaining consent from subjects, where the user agrees or disagrees to our use of their Facebook and academic data. For those who agree, the second step is the Facebook Authorization dialog, where the user authorizes our application to

access parts of their Facebook profile. We request the `read_stream` permission, which allows access to the user's News Feed. We show the consent form as a popup using Javascript, and the Facebook Authorization dialog is launched via a button generated by the Facebook Javascript SDK.

In order to verify the user's UCSD identity, we implemented UCSD Single Sign-On, which is a campus-wide username/password system. Once the user has verified their UCSD identity, we can link their Facebook ID and their UCSD ID, and have the authorization to gather the data necessary for our application. We collect the user's Facebook friend list using the Facebook API, and their course schedule through a SOAP web service provided by UCSD ACT (Administrative Computing and Telecommunications).

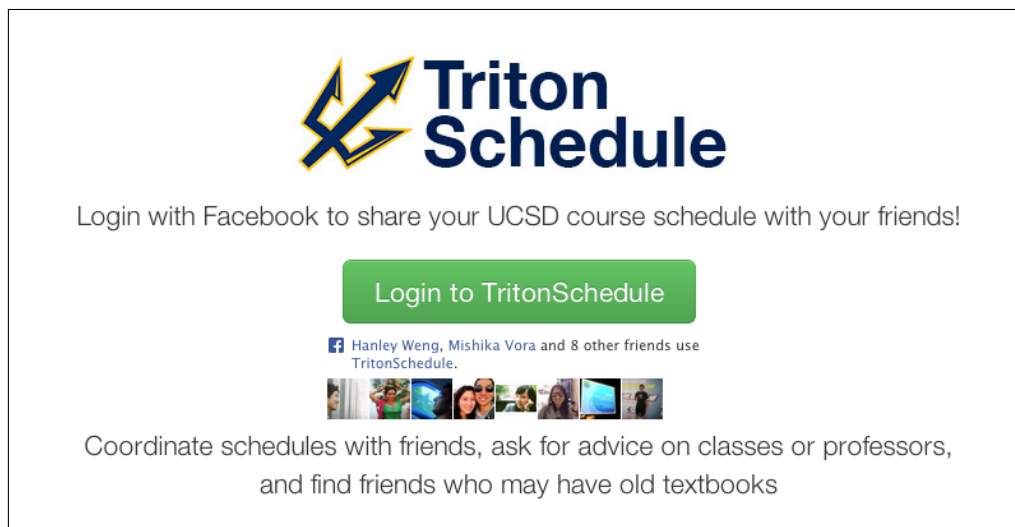


Figure 2.1: The TritonSchedule login page. The user sees pictures of their Facebook friends who are already using TritonSchedule.

The alternative to participation in this study is not to participate.

There may or may not be any direct benefit to you from participation. The investigator, however, may learn more about how social networks form, evolve, and influence student outcomes. This has the potential to benefit the university community at large.


Participation in research is entirely voluntary. You may refuse to participate or withdraw at any time without penalty by removing the TritonSchedule application from your Facebook account. Research will only be conducted using data from currently enrolled and consented individuals. Once you remove TritonSchedule from your Facebook account, we will stop recording your data.

For other questions or research-related problems, please contact James Fowler via email at jhfowler@ucsd.edu. You may call the Human Research Protections Program Office at (858) 657-5100 to inquire about your rights as a research subject or to report research-related problems.

By clicking on the Accept button below, you give your consent to be in this study.

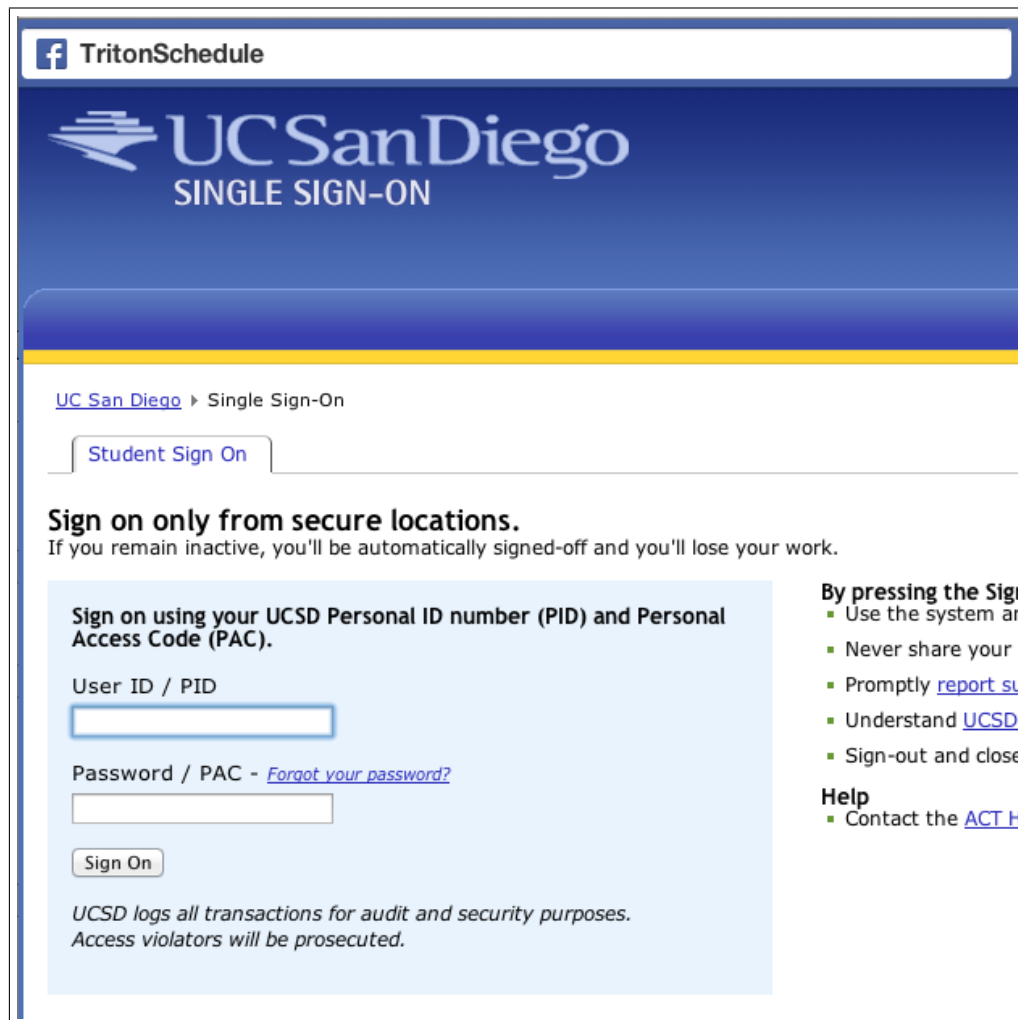
Figure 2.2: The consent page. This is shown before the Facebook Authorization dialog, so the user is required either to accept the terms of the study, or choose not to participate. See Appendix A for the complete consent form.

facebook Myoung Lah

 TritonSchedule would like to access your **public profile**, friend list, email address, News Feed, status updates, photos, likes and your friends' status updates, photos and likes.

[App Terms · Privacy Policy](#)

Figure 2.3: The Facebook Authorization dialog. This is a mechanism provided by Facebook for users to either authorize applications access to certain parts of their Facebook profile, or choose not to use the application. In the dialog, TritonSchedule asks for access to a user's News Feed.



f TritonSchedule

UC San Diego
SINGLE SIGN-ON

[UC San Diego](#) > Single Sign-On

Student Sign On

Sign on only from secure locations.
If you remain inactive, you'll be automatically signed-off and you'll lose your work.

Sign on using your UCSD Personal ID number (PID) and Personal Access Code (PAC).

User ID / PID

Password / PAC - [Forgot your password?](#)

Sign On

*UCSD logs all transactions for audit and security purposes.
Access violators will be prosecuted.*

By pressing the Sign On

- Use the system as intended
- Never share your |
- Promptly [report su](#)
- Understand [UCSD'](#)
- Sign-out and close

Help

- Contact the [ACT H](#)

Figure 2.4: The UCSD Single Sign-On page. Users must verify their UCSD identity before using TritonSchedule.

2.2.3 Features

TritonSchedule shows the user their current course schedule, as well as their complete course history. For each course, TritonSchedule shows all Facebook friends (who are also using TritonSchedule, and signed up for our study) who are currently enrolled in that course, or have taken that course in the past. This can be a useful tool for students to coordinate schedules, ask friends for advice on professors, and find friends who may have old textbooks.

The screenshot shows the TritonSchedule Facebook page. At the top is the Facebook search bar. Below it is the TritonSchedule logo and navigation tabs: 'My Schedule', 'All Courses', 'All Friends', and 'About'. The 'My Schedule' tab is selected. Below the tabs is a description: 'View your current and past course schedules, and connect with friends also enrolled in your courses'. There are 'Like' and 'Send' buttons. The main content is divided into two sections: 'Winter 2013' and 'Fall 2012'. Each section contains a table with 'Course Title' and 'Friends Enrolled' columns.

Course Title	Friends Enrolled
CSE 292 - Faculty Research Seminar	
CSE 221 - Operating Systems	
CSE 298 - Independent Study	
CSE 160 - Intro to Parallel Computing	

Course Title	Friends Enrolled
CSE 123 - Computer Networks	
MED 267 - Model Clinical Data Knowledge	
CSE 250A - Artificial Intel:Search&Reason	
CSE 298 - Independent Study	

Figure 2.5: The main TritonSchedule page. Students can see friends who are in the same courses, given that the friends have also signed up for TritonSchedule.

2.2.4 Benefits of our Study

As our study is planned to study student behavior over a long period of time, we would like to have ongoing access to participant’s Facebook data. The authorization to access Facebook data, an `access_token`, expires every two months, so it is important that users continue to use TritonSchedule and refresh the `access_token`. We believe that TritonSchedule provides a unique service that students will find useful throughout their time at UCSD, and is a good model for continued user engagement. Facebook applications are also easily shared with friends on Facebook, so we hope that this helps with user recruitment as well.

2.3 Gathering Facebook Data

While the Facebook application shows the News Feed as a collection of posts authored by a user’s friends, the `feed` endpoint of the Facebook Graph API returns only posts authored by the user. Posts authored by the user have likes and comments from friends, allowing us to count “incoming interactions.” This means that in order to count both “incoming” and “outgoing” interactions and make inferences about tie strength (see the Background section), we had to gather the Facebook News Feed of users as well as the News Feeds of their friends.

To gather the News Feed for a single user, we made requests to the `feed` endpoint of the Graph API using the `access_tokens` of study participants. The results from the Graph API are limited in size, and sometimes contain incomplete likes and comments information. To get complete News Feed data, we paged through the results, making multiple calls to the Graph API. To retrieve missing likes and comments, we made requests to the `likes` and `comments` endpoints of the Graph API.

Currently, the Graph API allows access to a user’s News Feed from the current time to the time they joined Facebook. This long-term data allows us to calculate tie strength during different times, and see how a user’s circle of close friends changes over time.

2.4 Gathering Academic Data

We connect directly to the ACT DB2 database server from our secure environment to gather academic data for students who are participating in our study. We will have access to the full academic record of students in our study, including class grades, withdrawal from a class, and major. These data will allow us to see how student performance changes over time, by calculating GPA per quarter, and by major and non-major classes. Other features we hope to gather are address, hometown, and UC Admissions Index, which is a composite of high school GPA and SAT or ACT scores [15]. These data will allow us to see how closely tie strength is correlated with geographic proximity, and if previous academic performance is related to current performance.

2.5 Technical Challenges

2.5.1 Facebook API Limit

Facebook places limits on how often requests are made to the Graph API. There are limits per application, per `access_token`, and per IP address. Generally, the limit is 600 API calls per 600 seconds (one per second). It takes on the order of 20 calls to the Graph API to gather the News Feed for a single user. Given that there a user may have anywhere from 500-1000 Facebook friends, and we have a goal of at least 1000 participants, we already have several orders of magnitude more API calls than are allowed within the limit.

In addition to the limit on the frequency of requests to the Graph API, there is a limit on the CPU utilization our application can cause. When we ran our data collection scripts and reached 600% of our Graph API frequency limit, our CPU utilization was over 8,000% of the allowed amount. It turns out that our use of the feed endpoint is very computationally expensive, and we need to either find a more efficient way of using the Graph API, or dramatically scale back on the amount of data we are collecting.

It is likely that in the future, we will have to settle for collecting only the

user's News Feed. This means that when inferring tie strength, we will only have information about incoming interactions (likes and comments that a user's friend has made on a user's posts). Given that Facebook activity is highly reciprocal, it is likely that the tie strength results will still be good.

2.5.2 Data Size

The Facebook News Feed data is returned as JSON objects, and stored as text in the database. Projecting from data collected so far, the average UCSD student has around 550 Facebook friends, meaning that we will need 1 TB or more to store complete News Feed data for 1000 users and their friends. We enabled compression (InnoDB in MySQL) on the table used to store News Feed data, which should help keep the size of the data down. Given that calculating the tie strength for each user requires parsing the news feed of each of their friends, it may become necessary to use distributed computing tools (like Apache Hadoop) to process the data in a timely manner. Currently, calculating tie strength for a user with 500 friends takes about 3 minutes, so calculating tie strength for 1000 users would take 50 hours. However, if we only gather the user's News Feed (due to the Facebook API limits), we will not face the roughly 1000x multiplication in data size caused by gathering all the friends' News Feeds, and data storage and processing time will no longer be an issue.

2.5.3 Facebook API Changes

The Facebook API is continually changing, and over the last two years, we had to make several changes to the application to keep it up to date. We expect that changes will be necessary to keep the application in compliance with the Facebook API, and to keep our study running.

Chapter 3

Results

3.1 General Statistics

TritonSchedule has been live since February 2013, and as of June 2013, 282 UCSD students signed up for the application. Of these students, 66 students are still active users, meaning they have accessed the application within the last two months. Of 282, 172 students are inactive, but they are still part of the study. The remaining 44 students have left the study by removing the application from Facebook.

We can infer that most TritonSchedule users are in the Computer Science and Engineering (CSE) department. Of the 282 students, there were 6173 enrollments among 1096 distinct courses. This means that on average, 6 students took the same course. However, the top 8 distinct courses account for 830 of the 6173 enrollments (13.4%), with an average of 104 students who took each course. These courses are MATH 20C, CSE 12, MATH 20B, CSE 15L, MATH 20F, CSE 21, CSE 20, and CSE 100, which are all part of the core CSE curriculum at UCSD.

TritonSchedule users have an average of 554 Facebook friends. As of 2012, Facebook users age 18-24 in the US (the demographic group most comparable to UCSD students) have an average of 429 Facebook friends [12]. The average Facebook user overall has 262 Facebook friends [12].

Considering only TritonSchedule users, students are connected to an average of two other students in the study. While this is quite low, there is a densely

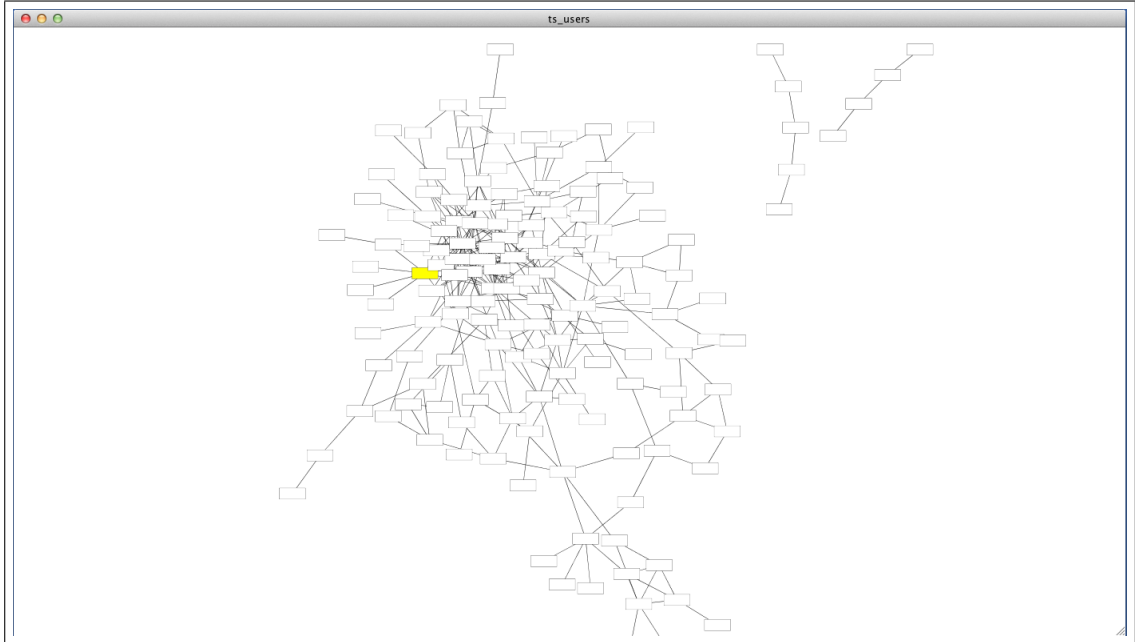


Figure 3.1: The largest connected component of the social network of around 180 TritonSchedule users. The author is highlighted in yellow.

connected subset of the study population as seen in Figure 3.1. This suggests that the application is spreading as friends recommend the app to each other, and that we should beware of biases in our study population (for example, the vast majority of current users are in CSE).

3.2 Inferring Tie Strength

As we are considering inferring tie strength using only incoming interaction information (through the user’s News Feed), we thought it would be useful to validate this method using complete interaction information for a small sample. We were able to gather complete data (user’s News Feed and all friends’ News Feeds) for the author and three other users. We first asked each user for a list of their closest friends. We then showed each user a list of their Facebook friends, ranked by the number of interactions. We generated two lists, one counting both incoming and outgoing interactions (as done in [13]), and one counting only incoming interactions (as we propose to do in the future). While this experiment was done

informally, and the sample was too small for any definitive evaluation, we do have several qualitative results and observations.

We were able to partially validate the tie strength calculation done by Jones et al. [13]. Because of the limits on the Facebook API, we were only able to gather complete user and friend News Feed data for four users (the author and three friends). For these users, we asked for a list of their 20 closest friends. We generated two lists from the News Feed data: 1) counting the “complete interactions” (outgoing and incoming) between the user and their friends, and 2) counting the “incoming interactions only.” In our sample, of the stated close friends, 3 to 15 of the top 20 friends were in the “complete interactions” list, and 4 to 13 of the top 20 friends were in the “incoming only” list.

While there was not a large quantitative difference in the number of stated close friends appearing on the two lists, there were several qualitative differences, where some users moved up or down significantly between the two lists. One explanation is that friendship is directional, and it is possible that while the user interacts a lot with the friend’s posts, the interaction is not reciprocated. Another explanation is that some friends authored a lot of posts, but did not like or comment a lot. This means that the user may have interacted with the friend (outgoing) much more than the friend interacted with the user (incoming). This could have caused some friends who were ranked highly on the “complete” list to drop (or not appear at all) on the “incoming only” list. The reverse is also possible. This suggests that a method of improving the tie strength results would be to normalize each friend’s interactions based on the frequency of their interactions. However, we are likely only able to collect incoming interaction data. While the two lists differed in the ranking of friends, the “incoming only” list had similar results to the “complete interactions” list, indicating that using incoming interactions only for calculating tie strength is a viable option. Again, these results are only for $n=4$.

There are some additional factors that affect the quality of our results. A perhaps obvious factor is that any friends who are not Facebook users cannot be ranked by this method. Additionally, a friend who is a Facebook user can

restrict their privacy settings so that their News Feed is not visible to applications, preventing us from collecting data on that friend. We can hypothesize that for our study population of UCSD students who are also Facebook users, a high proportion of their friends will also be Facebook users [7].

Chapter 4

Discussion and Future Work

Although we were not able to complete the compilation of our proposed dataset, we have laid the groundwork for future research on students and academic performance, hopefully in the near future. Namely, we have obtained IRB approval for our study, built partnerships with ACT in order to collect course schedule and academic data, and launched a Facebook application that provides a useful service to students and enables our study.

Our immediate concern is to work through our issues with the limits on the Facebook API, either by changing the way we use the API, or by collecting less data. We are also currently in touch with our contacts at Facebook to see if there are any other options available. We also want to grow the number of study participants by advertising to students where they are most likely to respond, and offering incentives to sign up. Once we have compiled the social network data with estimated tie strength, we can perform the clustering and causation analysis described in the Background section. Also, we are collaborating with researchers who are interested in combining the social network data with datasets on network security and health behaviors.

4.1 Future Work

4.1.1 Combination with Network Security

There is a group at UCSD that has compiled a data on network security behavior of machines on the UCSD network. Some measures included in this dataset are use of personal firewalls, anti-virus software, and network usage. If we are able to combine the network security data with the social network data, it is possible to examine clustering and causation for network security behavior much in the same way we have proposed for academic outcomes (see Background section).

4.1.2 Combination with Health Behaviors

The Student Wellness Office at UCSD is planning a survey to UCSD undergraduates that will ask them about their personal health behaviors. The Wellness Office has agreed to show a link to TritonSchedule at the end of this survey, and we have planned an incentive to students to sign up for our study. Given that by completing the health survey students have demonstrated their willingness to share their personal information for research, we hope that there will be a significant number of students willing to sign up for TritonSchedule as well. This will allow us to examine the health behaviors covered in the survey in the context of the social network.

4.1.3 Natural Language Processing

The Facebook News Feed we are gathering for each user is an array of posts, each post containing the text of the post, the users who have liked this post, and any comments left by users. We currently are focusing on counting likes, comments, and photo tags in order to infer tie strength. In the future, we can also use the text as input to Natural Language Processing (NLP) techniques to explore whether usage of certain words is predictive of particular behaviors, such as academic performance or health outcomes.

4.2 Lessons Learned

It took a large amount of administration and infrastructure cooperation to make the TritonSchedule project possible. One of the primary challenges was simply finding meeting times with university administrators who often had full schedules, requiring scheduling meetings up to a month in the future. Here are some additional considerations that may be helpful to others who would like to replicate this project at another institution.

4.2.1 IRB Review Process

Before submitting a study protocol to the IRB, we needed to develop our methods. We started developing a prototype of the Facebook application in June 2011. We had a meeting with the HRPP in July 2011 to discuss our project, and to keep ahead of any issues in our protocol. The study protocol was finalized and submitted in June 2012, we received conditional approval in September 2012, and final approval in October 2012. While it took almost 16 months from study conception to IRB approval, we were delayed significantly (between July 2011 and June 2012) in developing our prototype because of our request to access student records.

4.2.2 Accessing Student Data

We first met with the UCSD Registrar's Office from July 2011 to discuss our proposed project, and the possibility of accessing student course schedules to enable our Facebook application. There was significant resistance to allowing access to any student data because we were the first group they encountered that proposed combining academic and social network data in this way. Initially, we got a very low level of support, and work on accessing course schedules moved very slowly until we received final IRB approval in October 2012. After our study was approved by the IRB, we were able to receive full support from the Registrar's office, and our application was developed and launched to students by February 2013.

4.2.3 Development in a Secure Environment

As we were gathering sensitive student data, we were required to set up all of our application infrastructure in a HIPAA-compliant environment, managed by an IT department on campus. This managed infrastructure was definitely required to maintain compliance with HIPAA, as it would have been a huge burden to maintain the infrastructure ourselves. However, submitting tickets added some lead time to all infrastructure tasks.

Due to insufficient support from our first provider, we also faced the challenge of migrating our servers to a new provider. This required us to reconfigure parts of our application to suit the new security implementation. Additionally, there are generally fewer IT resources available in an academic environment, and there is more responsibility on developers to manage server configuration and security.

4.2.4 Advertisement and Growing the Study Population

We faced a chicken-and-egg problem with getting students to sign up to use TritonSchedule: when the value of the application is directly tied to the number of people using it, how do you get anyone to sign up in the first place? We hoped to start with a small group of “early adopters” through advertisement, and hopefully reach a tipping point where students would hear about TritonSchedule from their friends. Given this, we found student usage of TritonSchedule to be both encouraging and discouraging. At its peak, it was encouraging to see several hundred active users of the application, which is a significant number. However, it was discouraging to see new user recruitment halt after our initial advertisements.

We sent email advertisements to several student groups on campus, including the CSE undergraduate mailing list, and some fraternities. We asked some undergraduates that we knew to try the application and recommend it to their friends. This seemed to account for the first 100 users who signed up. The author went to several large undergraduate CSE classes and made an announcement about TritonSchedule and handed out flyers with a link to the application. This was not very effective, as we saw only a handful of students sign up after each class

announcement. Further emails to other mailing lists seemed to result in another 50 to 100 users. It seems like most mass emails and class announcements are being ignored by students.

We are currently planning to include a link to TritonSchedule at the end of a large health behavior survey that goes out to undergraduate students. We hope that since survey respondents are willing to share information about their health with researchers, these students will be willing to share their social network data as well. We are also planning a monetary incentive (raffle for prizes) that could be used in conjunction with the survey, or advertised separately to students. Students may also be more interested in signing up for TritonSchedule if they hear about it when they are registering for classes, which is when the application is most useful. We are continuing to look at successful social applications (like Twitter and Facebook itself) to learn more effective recruitment methods, and understand what motivates user adoption of new applications.

Appendix A

Informed Consent Form

University of California, San Diego
Assent to Act as a Research Subject

The TritonSchedule Study

James Fowler, Ph.D. is conducting a research study to find out more about social networks among UC San Diego students.

We expect to recruit at least 5000 users who will take advantage of a new Facebook application called TritonSchedule that allows students at UC San Diego to share class information with friends and to volunteer for this social network study. We will continuously collect data over the entire course of the project (at least 5 years).

In this study, we ask you to share two kinds of information:

1. Class Information

By clicking the agree button below, you are consenting to permit UC San Diego to access and share your class enrollment information with this Facebook application (TritonSchedule). Class enrollment data will be obtained from your TritonLink account and will include the course and section information for each course enrolled for the current and all previous quarters. Changes in your enrollment such as drops and adds will be reflected as they occur, but no grades or Withdrawals

will be shown. Class enrollment information will only be available within this application and only shared with others within this application that you have designated as friends on Facebook. This information is not available outside of the application and is not available to any of your friends who are not also users of this application. By clicking the agree button below you are consenting in accordance with Family Educational Rights and Privacy Act regulations and University policy to make your personal information available to others as defined by this disclosure statement.

2. Facebook Information

By clicking the agree button below, you are also consenting to permit the investigators to access information from your Facebook account, including your email, profile, news feed, timeline, and friend connections to other users, for the purpose of research. For the purpose of research, we may match this information to other sources of data about students held by UC San Diego, such as records held by the Registrar or Student Affairs. This includes data about institutional GPA, major, etc., but individually-identifiable data will be accessible only by research personnel and identifiers will be removed from the final data set.

There is a risk of loss of confidentiality, but the following steps will be taken to minimize this risk. Data will be stored on secure systems at iDASH [[link to: <http://idash.ucsd.edu/>]]. Identifiers will be used only for matching data, and will be stored in a separate, password-encoded file, available only to study personnel. To protect the confidentiality of user data, all records will be de-identified and reported in aggregate in any publications based on this study.

Refusal to participate in this study will not affect your academic standing with the University.

This study may involve risks that are currently unforeseeable. However, if any new risks become known in the future you will be informed of them.

The alternative to participation in this study is not to participate.

There may or may not be any direct benefit to you from participation. The investigator, however, may learn more about how social networks form, evolve, and influence student outcomes. This has the potential to benefit the university community at large.

Participation in research is entirely voluntary. You may refuse to participate or withdraw at any time without penalty by removing the TritonSchedule application from your Facebook account. Research will only be conducted using data from currently enrolled and consented individuals. Once you remove TritonSchedule from your Facebook account, we will stop recording your data.

For other questions or research-related problems, please contact James Fowler via email at jhfowler@ucsd.edu. You may call the Human Research Protections Program Office at (858) 657-5100 to inquire about your rights as a research subject or to report research-related problems.

By clicking on the Accept button below, you give your consent to be in this study.

Bibliography

- [1] Patrick, Kevin, Jennifer R Covin, Mark Fulop, Karen Calfas and Chris Lovato. Health Risk Behaviors among California College Students. *Journal of American College Health*, 1997; 45:6, 265-272.
- [2] McPherson, Miller, Lynn Smith-Lovin and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 2001; 27:415-44.
- [3] Doshi, Amol, Kevin Patrick, James F Sallis and Karen Calfas. Evaluation of Physical Activity Web Sites for Use of Behavior Change Theories. *Annals of Behavioral Medicine*, 2003; 25(2):105111.
- [4] Christakis, Nicholas A and James H Fowler. The Spread of Obesity in a Large Social Network over 32 Years. *NEJM*, 2007; 357:370-9.
- [5] Christakis, Nicholas A and James H Fowler. The Collective Dynamics of Smoking in a Large Social Network. *NEJM*, 2008; 358:2249-58.
- [6] Fowler, James H and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ*, 2008; 337:a2338.
- [7] Lewis, Kevin, Jason Kaufman, Marco Gonzalez, Andreas Wimmer and Nicholas Christakis. Tastes, Ties, and Time: A New Social Network Dataset Using Facebook.com. *Social Networks*, 2008; 30:330-342.
- [8] Gilbert, Eric, and Karrie Karahalios. Predicting tie strength with social media. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2009; pp. 211-220.
- [9] Kahanda, Indika and Jennifer Neville. Using Transactional Information to Predict Link Strength in Online Social Networks. *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM'09)*, 2009; pp. 74-81.
- [10] Munson, Sean A, Debra Lauterbach, Mark W Newman and Paul Resnick. Happier Together: Integrating a Wellness Application Into a Social Network Site. *Proceedings of the Persuasive 2010*, Springer, 2010; 27-39.

- [11] Newman, Mark W, Debra Lauterbach, Sean A Munson, Paul Resnick and Margaret E Morris. "It's not that I don't have problems, I'm just not putting them on Facebook": Challenges and Opportunities in Using Online Social Networks for Health. *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, 2011.
- [12] Edison Research, Arbitron. (June 2012). The Social Habit. Retrieved June 13, 2013, from <http://socialhabit.com/secure/wp-content/uploads/2012/07/The-Social-Habit-2012-by-Edison-Research.pdf>.
- [13] Jones, Jason J, Jaime E Settle, Robert M Bond, Christopher J Fariss, Cameron Marlow, James H Fowler. Inferring Tie Strength from Online Directed Behavior. *PLoS ONE*, 2013; 8(1):e52168.
- [14] US Department of Health and Human Services. (n.d.). *Summary of the HIPAA Security Rule*. Retrieved June 13, 2013, from <http://www.hhs.gov/ocr/privacy/hipaa/understanding/srsummary.html>.
- [15] University of California. (n.d.). Statewide path. In *University of California Admissions*. Retrieved June 13, 2013, from <http://admission.universityofcalifornia.edu/freshman/california-residents/admissions-index/index.html>.
- [16] University of California, San Diego. (n.d.). *UCSD Human Research Protections Program*. Retrieved June 13, 2013, from <http://irb.ucsd.edu/>.