Towards the Construction of Next Generation Online Discussion Boards – Mining Online Conversations at Channel Islands for Useful Design Patterns

Research Mentor && Student Researchers

Dr. Brian Thoms, Asst. Professor of Computer Science and Information Technology
Glo Mercado, Student in Computer Science and Mathematics at CSUCI
Matthew Wells, Student in Sociology at CSUCI

Project Background

Engagement Theory asserts that students must be meaningfully engaged in learning activities through interaction with others, facilitated and enabled by technology [1]. One such activity, which has become an integral component across Learning Management Systems (LMS) is the asynchronous online discussions (AOD). For many online and hybrid courses, the AOD is an activity that helps bind an individual's learning experience to the course community. AODs can be conceptualized by their ability to facilitate cognitive, on-topic, on-task, and sustained discussion among students [2]. However, a problem with existing AODs is that they fall short in fostering the levels of interaction seen in online social networking (OSN) software [3]. Maurino [4] found that while students may not be entirely happy with the use of threaded discussions, they did desire student interaction and collaboration.

Depicted in Figure 1, this research project explores the implementation of a sentiment analyzing discussion-board (S.A.D.) as a mechanism to foster greater levels of interaction and keep students engaged in course content. Kadushin [5] asserts that interactions lead to sentiments, which can be positive or negative, but positive sentiments lead to further interaction and negative sentiments lead to less interaction. The goal of S.A.D. was to encourage and elicit more positive discussion posts with the goal of fostering greater interactions.

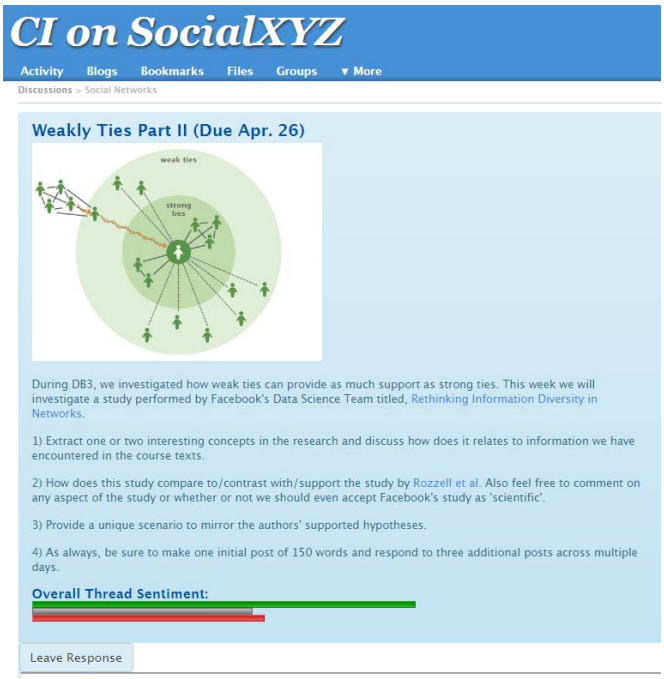

Figure 1 - S.A.D. System

To measure the impact of S.A.D., social networking data for two courses was analyzed. Over 2500 online conversations, consisting of discussion data, blogs and tweets, were mined for sentiment (positive, negative or neutral) through the Natural Language Toolkit (NLTK) application programming interface (API). NLTK is a broad-coverage natural language toolkit that provides a simple, extensible, uniform framework natural language processing [6] and incorporated into the system design. Additionally, SNA graphs were constructed in NodeXL, which is an open source extension for Microsoft Excel that provides a range of basic network analytics and visualization features [7].

Connection to CI Mission(s)

This project addressed multiple mission areas including 1) undergraduate research, which allowed for two undergraduate researchers, 2) collaborative learning, where the primary focus of the research was to analyze software designed to engage students taking interdisciplinary courses and 3) online learning communities, which, as my prior work in this space can attest, can be facilitated by innovative internet-based social networking software. Also in line with the mission of the university, this research project adopted an interdisciplinary research approach, which brought students from Sociology and Computer Science to collaborate on the issue of student engagement through the lens of a socio-technical framework.

Spring 2016 Success

From January through May, I had the opportunity to work with two very talented undergraduate students, Glo Mercado, a dual major in Computer Science and Math and Matthew Wells, a major in Sociology.

During this timeframe, we were tasked with continuing research started during Project Acceso in Summer 2015. Photo 1 showcases the original research team with Glo, second from the right, and Dr.

Thoms, first from the left. During Project Acceso, our research team analyzed discussion board data from COMP 342, Internet-based Social Networking. The results led to the design of S.A.D., a sentiment analyzing discussion-board, which was implemented for COMP 342 in spring 2016. An initial analysis and design of this project has been prepared for submission to the Hawaiian International Conference on System Sciences this June.

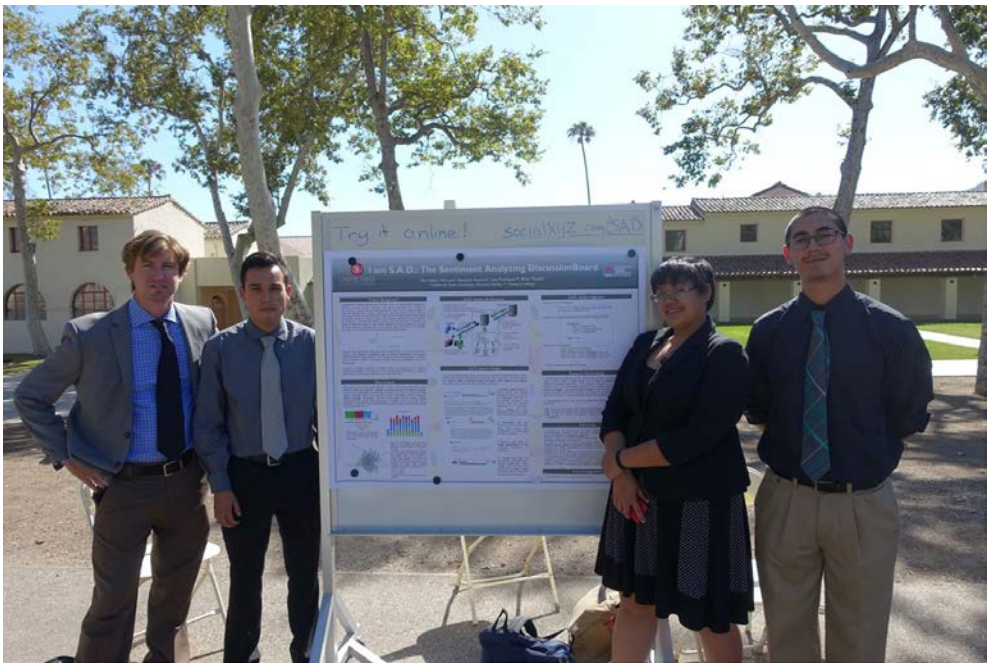

Photo 1 - 2015 Acceso Team (Left to Right: Brian Thoms, (Advisor), Benjamin Ramirez (Student Researcher), Glo Mercado (Student Researcher), Jose Rodriguez (Student Researcher))

Continuing this research during the 2016 spring semester Glo and Matthew (shown working diligently in Photo 2) helped analyze two semesters worth of social networking data, mining this data for trends and patterns that will be integrated into the analysis portion of an upcoming academic research paper. The analysis allowed Glo and Matthew to adopt popular techniques in data mining and social network analysis.

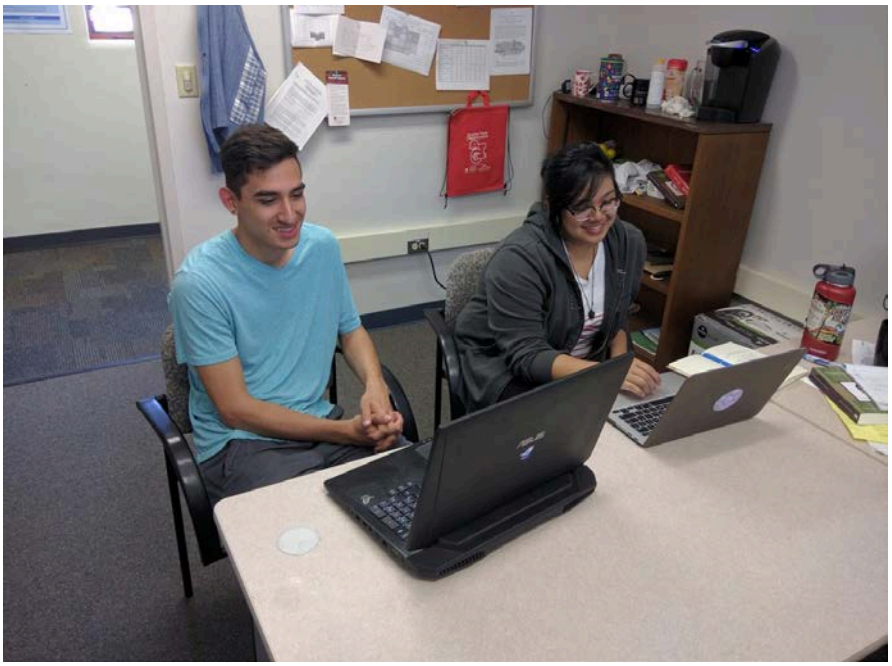

Photo 2 - Matthew Wells and Glo Mercado working hard in the lab.

The data Matthew and Glo analyzed specifically looked at how sentiment evolves and patterns emerge as a discussion board grows and ages (i.e. more students participate and fluctuations across the semester). As briefly touched upon in the introduction, extant literature asserts that positive interactions produce more interactions, which in turn produces more positive interactions. While the premise is simple, developing a system that aims to promote such positive interactions is not, especially as it relates to "homework".

Project Outcome(s)

Phase I

Phase I of our spring tasks focused on a literature review, specifically literature related to sentiment and online discussion board interactions. Below is a listing of just a handful of extant academic research articles Glo and Matthew unearthed this semester in the area of online discussion boards, sentiment analysis and online learning. This part of the research process, while not the most glamourous or fun, is an important step for understanding what research, if any, is on-going across these spaces. It was encouraging to discover that very little research has explored how sentiment influences learning participation in online learning communities. Some additional papers we discovered include:

1. Croxton, RA (2014) 'The role of interactivity in student satisfaction and persistence in online learning', Merlot Journal of Online Learning and Teaching, 10(2), pp: 314-325.
2. Garje GV, Inamdar, A, Mahajan, H, Bhansali, A and Khan, SA (2016). "Stress Detection and Sentiment Prediction: A Survey," International Journal of Engineering Applied Sciences and Technology, 1(2), pp: 32-34.
3. Kumar A and Sebastian, TM (2012). "Sentiment Analysis: A Perspective on its Past, Present and Future," International Journal on Intelligent Systems and Applications, v10, pp: 1-14.
4. Mihail, R, Rubin, B and Goldsmith, J (2014). "Online Discussions: Improving Education in CS?," SIGCSE'14, March 5-8, Atlanta, GA, USA

5. Murthy, D and Petto, LR (2015). “Comparing Print Coverage and Tweets in Elections: A Case Study of the 2011-2012 U.S. Republican Primaries,” Social Science Computer Review. 33(3), pp: 298-314.
6. Ortigosa, A, Martin, JM and Carro, RM (2014). "Sentiment analysis in Facebook and its application to e-learning," Computers in Human Behavior, v31, pp: 527–541.
7. Ranco, G, Aleksovski, D, Caldarelli, G, Grcar, M and Mozeti, I (2015). "The Effects of Twitter Sentiment on Stock Price Returns" PLoS One, 10(9).
8. Wong,JS, Pursel, B, Divinsky, A and Jansen, B (2015). "Analyzing MOOC Discussion Forum Messages to Identify Cognitive Learning Information Exchanges," ASIST 2015, November 6-10, St. Louis, MO.
9. Wen, M, Yang, D and Rose, CP (2014). "Sentiment Analysis in MOOC Discussion Forums: What does it tell us?," Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014), July 4, 2014 - July 7, London, UK.
10. Zarra, T, Chiheb, R, Faizi, R, and El Afia, A (2016). "Using Textual Similarity and Sentiment Analysis in Discussions Forums to Enhance Learning," International Journal of Software Engineering and Its Applications, 10(1), pp: 191-20;

Phase II

Part II of our project re-focused attention on the data, specifically discussion data from 2015 and 2016 spring COMP 342 courses. The analysis phase can be broken down into two distinct parts, 1) social network analysis and 2) sentiment analysis.

1. Social Network Analysis

Detailed in Tables 1 and 2 and Figures 2 and 3, Glo and Matthew tackled an SNA analysis between 2015 and 2016 discussion data. Focusing attention on one factor, density, we can learn a lot. Density is often measured to be the heart of a social network and is used to determine the strength of the ties between all individuals in that network. Alone, this number is baseless, but when compared against a benchmark, the number can mean a great deal. Consequently, when we compared 2015 and 2016 COMP 342 data, we discovered that students participating in spring 2016 maintained a more dense social network than those from 2015. This was surprising for the simple reason that as a network become larger (18% in the case of 2016), density generally decreases (think Facebook, or even the Universe, for example). In an educational setting, this often holds true and as a classroom population grows, meaning more students are enrolled, the less likely all students will be able to connect with more students. Although more analysis is still required, our goal is to attribute S.A.D.’s influence on these increased interactions.

2015 Social Network Analysis

Graph Metric	Value
Graph Type	Directed
Vertices	19
Unique Edges	184
Edges With Duplicates	0
Total Edges	184
Self-Loops	2
Reciprocated Vertex Pair Rati	0.48
Reciprocated Edge Ratio	0.65
Maximum Geodesic Distance	2
Average Geodesic Distance	1.21
Graph Density	0.53
Modularity	NA
NodeXL Version	1.0.1.350

Table 1 - Spring 2015 Metrics

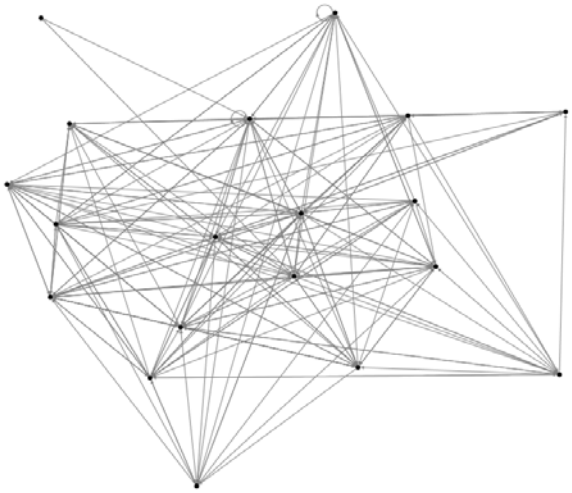

Figure 2 - Spring 2015 SNA Graph

2016 Social Network Analysis

Graph Metric	Value
Graph Type	Directed
Vertices	22
Unique Edges	271
Edges With Duplicates	0
Total Edges	271
Self-Loops	3
Reciprocated Vertex Pair Rati	0.47
Reciprocated Edge Ratio	0.64
Maximum Geodesic Distance	2
Average Geodesic Distance	1.16
Graph Density	0.58
Modularity	NA
NodeXL Version	1.0.1.350

Table 2 - Spring 2016 Metrics

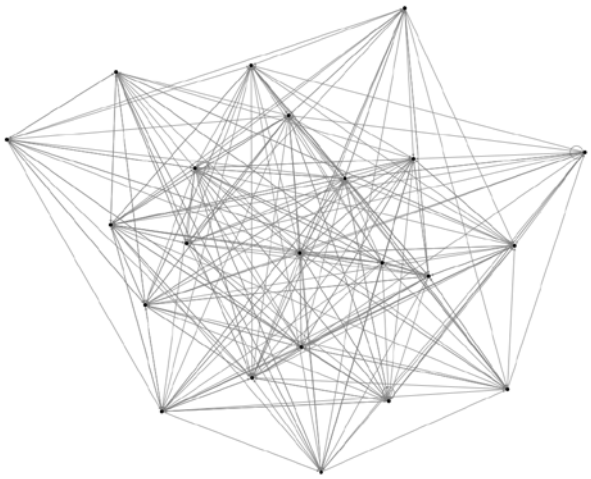

Figure 3 - Spring 2016 SNA Graph

2. Sentiment Analysis

During Phase II we also performed a rigorous sentiment analysis, which meant that Glo and Matthew painstakingly analyzed thousands of social media data, mining the data for sentiment and uncovering hidden patterns that existed in the data. Detailed in Table 3 and Figure 4 are the high-level summaries of discussion data across 2015 and 2016.

	Positive	Pos %	Negative	Neg %	Neutral	Neu %	Total Posts	Avg. Per Student
Sp2015	279	50%	174	31%	110	20%	563	29.6
Sp2016	351	50%	175	25%	175	25%	701	30.5

Table 3 - Sentiment Analysis 1

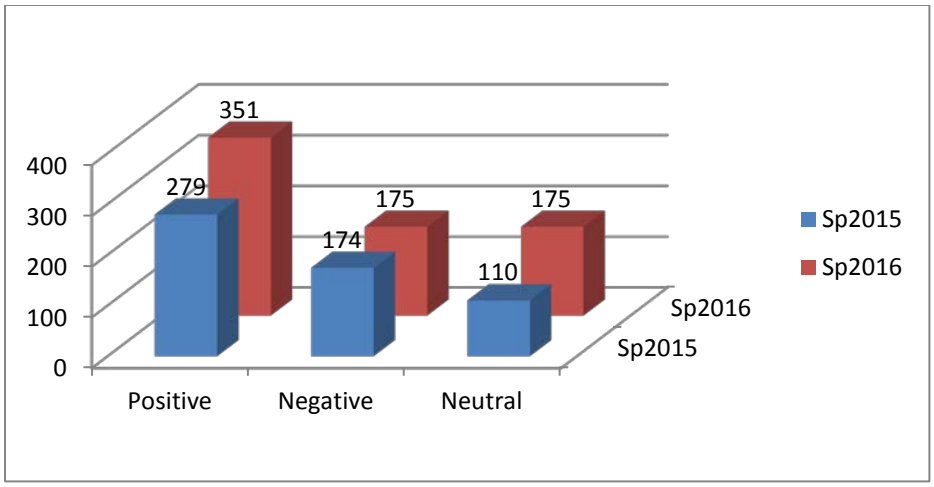

Figure 4 - Sentiment Analysis 2

This data only provides a starting point for analyzing the interaction patterns across the social network.

Dividing the tasks, Glo focused her efforts on the macro-level discussion, or the discussion boards as a whole. Specifically, she was tasked with analyzing the number of posts, versus the overall sentiment of the discussion board. In a nutshell, she determined that our hypothesis was correct and the more posts a discussion board has, the more positive the sentiment.

Matthew focused efforts on the micro-level discussion, or an individual's discussion board footprint. One question we posed was whether individuals who make more positive posts, receive more positive responses. What we found was the contrary. It seems that response posts to positive posts do not match, on average, the index of the owner's post.

Phase III

The goal of Phase III was to redesign S.A.D. to introduce new ways sentiment could be enhanced to better engage student interactions. Phase III was always ambitious and required more resources than currently available. And it is at the phase I plan to continue working during Acceso 2016.

Project Evaluation and Success

Overall, I would classify this project as a success for two primary reasons. First and foremost, I was able to bring two students from different disciplines together, introduce them to my research in social informatics and human computer interaction and help them to analyze and interpret large quantities of social data. Secondly, this project allowed me to further my own research agenda as I continue to investigate innovative software design in online social networking software. Working closely with Glo and Matthew allowed me to think more critically about my overall research objectives.

From the Research Assistants

At the end of the semester, I requested that each student researcher provide a short reflection of their experience working on this project. Both students provided a two-page reflection on how they approached this research project and what knowledge and/or skills they developed from their experience. Figures 5 and 6 provide a high-level tag cloud of their papers. A tag-cloud is a simple visual

depiction of user-generated content, which uses color and font size to represent the prominence or frequency of the words depicted.

Glo Mercado wrote:

"Another analysis we wanted to perform was the effect of thread length on sentiment. I knew the algorithm I would use to go about it, but I didn't know how to execute it in excel without having to do it by hand, so I looked for other ways to approach this. I had used Python for an assignment outside of this research, and we were able to import excel files using the Pandas library. In Python, I was able to implement the algorithm to get the threads, and even prepare the posts to get the overall sentiment values and readings of those posts."

Figure 5 – Glo's Reflection Tag-Cloud

Matthew Wells wrote:

"One of the last programs that I learned over the course of the research was NodeXL. NodeXL was a program that I had never even heard of before. This program was a challenge for me because there was a lot that I did not understand at first. However Professor Thoms was great at taking his time to explain how things worked with the program and where to place certain data. Once again, gaining the knowledge of how various programs was very helpful for the fact that all of these programs can have an application outside of this specific research."

Figure 6 - Matthew's Reflection Tag-Cloud

More Fun Facts!

Also during the past semester, both Glo and Matthew had the chance to bring this research on the road. Glo presented preliminary findings at the Women in Computing in Southern California Conference,

which took place on April 9-10 in Ventura, CA (http://cwicsocal16.calpoly.edu/program.html). During the same weekend, Matthew was able to visit the Facebook campus in Palo Alto, where he and a select group of students received a hands-on tour of where the big guns analyze petabytes worth of social networking data. Shown in Photo 3 (Matthew is fourth from the right and Dr. Thoms is first from the right), the trip was made possible through Institutional research funds, IRA #690 - Logging-on to Internet-based Social Networks Offline. Photos from this trip can be found online through the following link: https://goo.gl/photos/faCDrxn7mkHoFPcv7.

Photo 3 - Trip to Palo Alto

Photos 4 and 5 are a couple of additional photos of Glo and Matthew working crunching data!

Photo 4 - Glo Mercado pushing data to the edges (literally!)

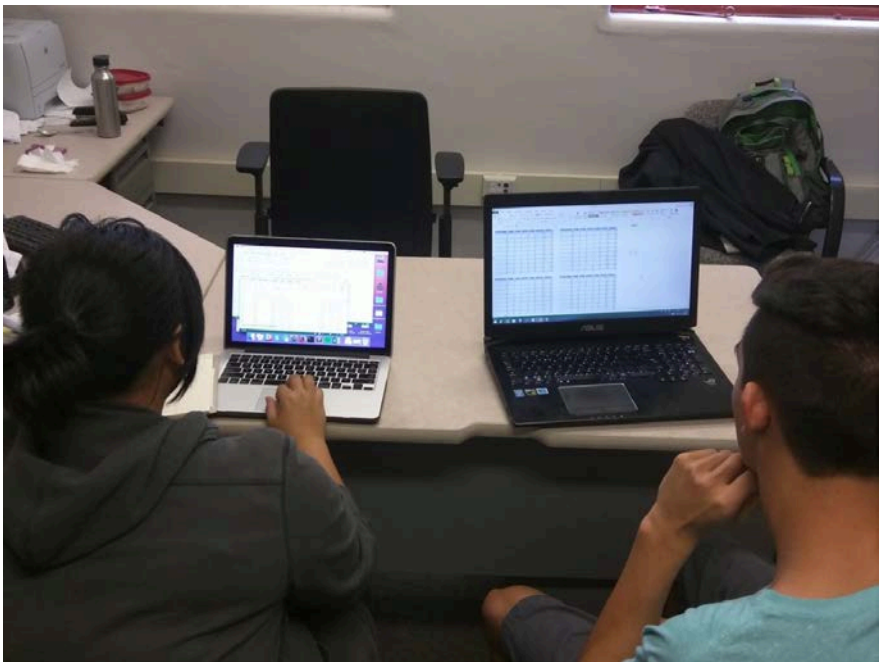

Photo 5 - Glo and Matthew at their workstations. Mac and Windows Friendly!

Acknowledgements

This research is largely dependent on various open-source software platforms. NodeXL Basic is a free, open-source template for Microsoft® Excel® and makes it easy to explore social network graphs [7]. With NodeXL, Glo and Matthew were able to enter network edges (i.e. discussion board data, blog data and twitter data) into a worksheet and at the click a button, having a starting point for a more deeper social network analysis, which was performed using resources from another open source tool, the

Natural Language Tool-Kit (NLTK). NLTK is a broad-coverage natural language toolkit that provides a simple, extensible, uniform framework natural language processing [6]. Figure 7 is a snapshot of what Glo and Matthew would encounter on a weekly basis.

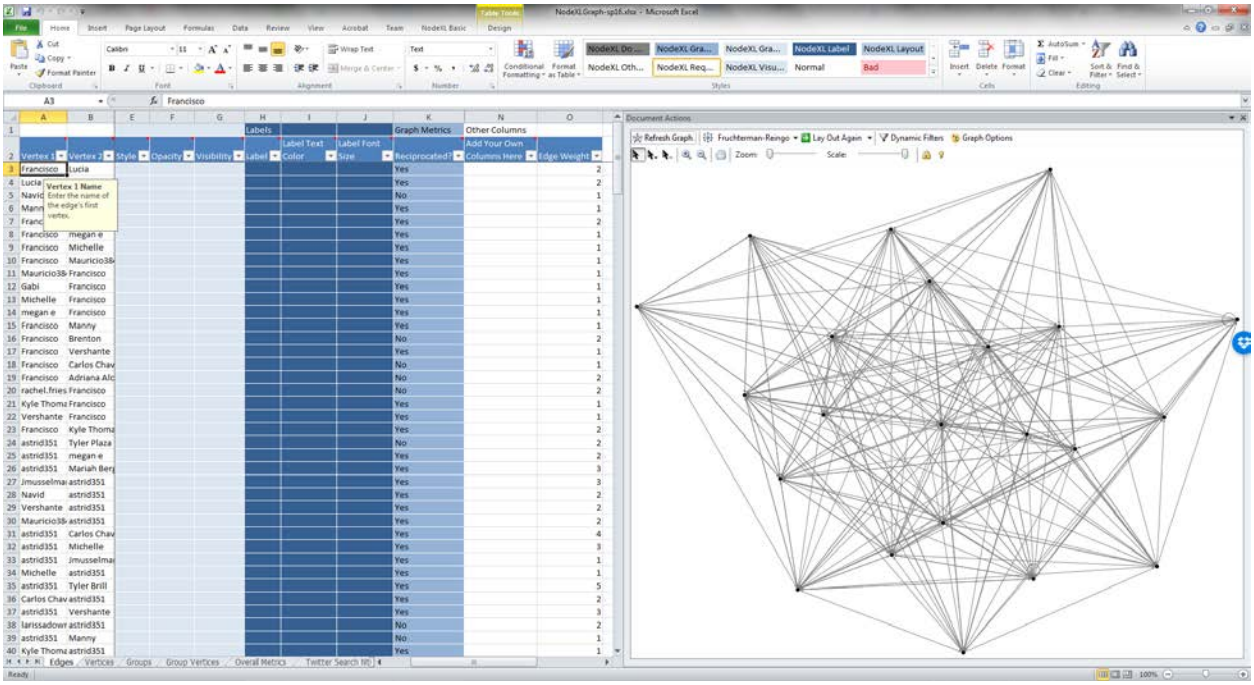

Figure 7 - NodeXL Snapshot

References

[1] Kearsley, G. and Schneiderman, B. (1999). “Engagement theory: A framework for technology-based learning and teaching,” Retrieved June 26, 2013, from http://home.sprynet.com/~gkearsley/engage.htm

[2] Pituch, K. and Lee, Y. (2006). “The influence of system characteristics on e-learning use,” Computers & Education, 47(2), 222-244.

[3] Thoms, B., Eryilmaz, E. (2014). “How Media Choice Affects Learner Interactions in Distance Learning Classes,” Computers & Education, v75, pp.112-126.

[4] Maurino, P. (2006). “Participation and interaction: F2F vs. online,” Academic Exchange Quarterly, 36(1), pp 129–143.

[5] Kadushin, C. (2005), "Networks and Small Groups," Structure and Dynamics, 1(1).

[6] Bird, B., Klein, E., Loper, E. and Baldridge, J. (2008). “Multidisciplinary instruction with the Natural Language Toolkit,” Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics, Columbus, Ohio, USA, June 2008.

[7] Hansen, D., Shneiderman, B. and Smith, M. (2011). Analyzing Social Media Networks with NodeXL: Insights from a Connected World, Burlington: Morgan Kaufmann.