Do You Know the Way to SNA?: A Process Model for Analyzing and Visualizing Social Media Data

Derek L. Hansen*, Dana Rotman*, Elizabeth Bonsignore*, Nataša Milić-Frayling†, Eduarda Mendes Rodrigues†, Marc Smith‡, Ben Shneiderman*, Tony Capone†

*University of Maryland, Human Computer Interaction Lab; †Microsoft Research; ‡Connected Action

ABSTRACT

Traces of activity left by social media users can shed light on individual behavior, social relationships, and community efficacy. Tools and processes to analyze social traces are essential for enabling practitioners to study and nurture meaningful and sustainable social interaction. Yet such tools and processes remain in their infancy. We conducted a study of 15 graduate students who were learning to apply Social Network Analysis (SNA) to data from online communities. Based on close observations of their emergent practices, we derived the Network Analysis and Visualization (NAV) process model and identified stages where intervention from peers, experts, and an SNA tool were most useful. We show how the NAV model informs the design of SNA tools and services, education practices, and support for social media practitioners.

Author Keywords

Social network analysis, visualization, social media, process model, NodeXL, online communities.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Social media services, such as Facebook, Twitter, Digg, among others, have enabled new forms of collaboration and interaction in nearly every imaginable human endeavor. And we have only begun to realize the potential of technology-mediated social interaction. Despite numerous success stories, we must remember the countless failures due to social and technical factors. How can we support practitioners in their efforts to cultivate meaningful and sustainable online interaction?

One promising strategy is to provide tools and concepts that help practitioners make sense of social media data. There is precedence to this approach in the development of sophisticated, yet fairly intuitive *website analytics tools* such as Google Analytics [12]. These tools help non-programmers understand website traffic patterns so they can make more informed design decisions. We envision an equivalent set of *social analytics tools* (e.g., [17, 21]) to help social media analysts and community administrators make better decisions based on their in-depth understanding of social participation and relationships. Social analytics, which includes Social Network Analysis (SNA), extends already complex graph analysis metrics and visualizations with exploratory data analysis approaches, and requires the engagement of professionals experienced in social interactions and social media contexts.

To gain acceptance by a broad range of practitioners, tools that reduce the complexity of data processing are vital. Eliminating the need to program custom algorithms for common processing tasks can make SNA more accessible. Moreover, enabling interactive visual exploration of data via a variety of layouts can aid in the discovery, understanding, and presentation of network properties. To varying degrees, several SNA toolsets such as UCINET, Pajek, SocialAction, and NodeXL have advanced toward these goals. However, as with any new practice, success depends on the common language and best practices that evolve among practitioners as they apply the tools in various scenarios and share their experiences and expertise. We need to understand and capture the processes that emerge as users explore social interaction to enhance their power to make sense of and manage interaction patterns.

With that in mind, we conducted a qualitative user study of graduate students learning to apply SNA concepts and tools to better understand online communities of their choice. We make two primary contributions. First, we derive the *Network Analysis and Visualization (NAV) process model* that emerged from the collective experience of students learning to use SNA metrics and visualizations. We identified stages within the model where interventions from peers, experts, and analysis tools are most useful. Second, we offer recommendations for making SNA tools and services more accessible to practitioners, especially novices. These include recommendations for (1) designers of SNA tools, (2) educators introducing SNA concepts to online community analysts, and (3) practitioners struggling to make sense of social media data. We found the fine-level granularity of the NAV process model invaluable when developing these recommendations – far more helpful than more generic sensemaking models, although the NAV model shared some high-level similarities with them.

RELATED LITERATURE

SNA and its mathematical companion, the graph theory, have a long and distinguished history of academic contributions [3, 11, 23]. Recently, many researchers have used SNA to examine social interaction in computer-mediated environments, helping to identify unique social roles [26], social structures [6], and dissemination patterns [2]. Despite SNA's success in academic circles and its appearance in mainstream publications [3] and management literature [9], it has not been widely used by practitioners. This is likely to change as more usable SNA tools are developed and as the historically onerous process of network data collection is replaced by automatic data collection from social media sources.

Process models that describe key activities, tasks, cognitions, and/or feelings have been useful in helping design novel tools [19] and educational interventions [13]. They are particularly good at identifying moments where interventions from peers, experts, or computational aids are most useful. Pirolli & Card [19] call these moments "leverage points" in their process model of information analyst's activities. Their work is part of a larger effort to characterize the *sensemaking process* of expert intelligence analysts [16, 22]. In a different, but related domain, Kuhlthau [15] developed a process model of information seeking behavior and identified stages where information mediators, i.e., educators, can help students the most.

Motivated by the success of these approaches we investigate the sensemaking process that emerges when both the SNA concepts and SNA tools are introduced as the means of data analysis. Our work extends existing literature on sensemaking models [16, 19, 22] on two fronts: (1) we observe novices, not experts, and (2) we focus on the social network analysis and visualization, tasks that have not been explicitly investigated from the sensemaking perspective.

Considering SNA toolsets, the development of SocialAction [18] was based on a Systematic Yet Flexible (SYF) framework that extended successful process models such as Amazon's checkout, TurboTax's income tax preparation, and the Spotfire Guides for visual analytics. The SYF framework organized network analysis into 7 steps: (1) overall network metrics (2) node rankings, (3) edge rankings, (4) node rankings in pairs, e.g., degree vs. centrality, plotted on a scattergram, etc., (5) edge rankings in pairs, (6) cohesive subgroups, e.g., finding communities, and (7) multiplexity, e.g., analyzing comparisons between different edge types, such as friends vs. enemies. These steps frame the process that experts – not students or novices – follow when exploring complex data sets.

Finally, the rapidly growing literature about information visualization often examines systems that support network visualization but in-depth user studies such as [18] and ours are rare. Heer and boyd [14] demonstrated that novices enjoy browsing data from social networking sites like Friendster. A survey of 77 researchers, mostly social scientists, showed a preference towards menu-driven general purpose packages, such as UCINET and Pajek, over programmable systems such as JUNG, GUESS, and Mathematica [1]. However, users expressed significant frustrations with all the systems due to challenges of learning complex interfaces [1]. The success of ManyEyes, a collaborative system for creating and sharing information visualizations, suggests the desire of many to make information visualization more accessible [25].

Our detailed user study using an SNA toolset equipped with a robust set of graph layout options can help characterize the process novices follow when analyzing network data with the aid of visualization tools, and offers insights useful to designers, educators, and community analysts hoping to broaden the adoption of SNA toolsets and expand a collaborative SNA knowledge base.

METHODS

Study Setup

We conducted a month-long user study of 15 students in a graduate course on Computer-mediated Communities of Practice (CoP).

Teaching Context

The CoP course is an elective, drawing graduate students from library science and information management. The purpose of the course is to help students become proficient community analysts, able to identify and apply appropriate technologies and social practices to help cultivate communities. The course includes a weekly classroom session and a website where students post weekly to a discussion forum and periodically to individual blogs.

The study took place during a 3-week SNA module that introduced SNA concepts and the NodeXL SNA tool, and their application to the data from online communities. The module occurred 1/3rd of the way through the CoP course and required students to analyze a community they had chosen to study throughout the semester.

The SNA module included a 2.5 hour, hands-on lab session that used the *Network Analysis with NodeXL: Learning by Doing* tutorial [13]. The tutorial followed a task-based framework of 9 steps: (1) basic, (2) layout, (3) visual design, (4) labeling, (5) filtering, (6) grouping, (7) graph metrics, (8) clustering, and (9) advanced. Course readings and discussions covered SNA and community metrics, social roles, and network visualization quality as measured by NetViz Nirvana guidelines about the network layout [10]: (1) every node is visible, (2) the degree of every node can be counted, (3) every edge can be followed from source to destination, and (4) clusters and outliers are identifiable.

NodeXL SNA Tool

NodeXL is a plug-in for Excel 2007 that exploits a widely used spreadsheet paradigm to provide a range of basic network analysis and visualization features [20]. The NodeXL template is a highly structured workbook with multiple worksheets to store information that is needed to represent a network graph. NodeXL visualization features allow users to display network representations using various layouts, apply filters, and map attributes of the nodes and edges to visual properties, including shape, color, size, transparency, and location. NodeXL was selected for teaching SNA because of its perceived ease of use and a broad coverage of SNA metrics and visualization features. Current versions and community support are available at: http://www.codeplex.com/nodexl.

Data Collection

Classroom Observations. Two researchers passively observed classroom instructions and discussions during the SNA module, creating video-recordings and taking detailed notes about student questions and comments.

Course Work. Student assignments included (a) a take-home quiz in which students used NodeXL to answer questions about a network derived from a company's internal discussion forum, (b) intermediate SNA visualizations and descriptions posted to the course website, and (c) final assignments that included two or more network visualizations and a one-page description of the visualization analysis. Students could focus on any aspect of their community and use any type of network visualization and analysis. They were graded on the quality of the network visualizations, following the "NetViz Nirvana" guidelines [10], the accuracy and the quality of the SNA description, and the importance of their analysis for understanding the community. Students were encouraged to help one another. Two voluntary lab sessions were set up to facilitate peer support.

Surveys & Group Modeling. Prior to the SNA module, students completed an online survey that assessed their familiarity with Excel, SNA, and social media. A hand-written, post-study survey was completed on the day when students turned in their final SNA assignment. During that session one of the researchers mediated a collective exercise in which students reflected upon the process they followed when applying SNA to social media data. Each student individually identified and mapped out the major activities and sub-tasks they engaged in during the SNA assignment. The researcher then facilitated the collaborative creation of a process model through group discussion and drawing on a chalkboard. The derived model incorporated most of the distinct stages from individual maps. Using it as a reference point, the researcher asked students to indicate stages and tasks they found most challenging and those where peer and expert advice was most helpful.

Diaries. While working on the quiz and the SNA assignments, students completed 3 semi-structured diaries. Self-reporting diaries have been used to inform process models before [15], allowing participants to record their actions and emotions with little interference from researchers [7]. Our diary form included open-ended questions about the assignments, the role of peer and expert support, resources that were used, and the satisfaction students had with their own work. Students also recorded the tasks they performed, time spent on tasks, and their feelings during tasks.

Individual Observations and Interviews. All but 2 students were observed and interviewed by one of two researchers. Eleven students were observed while completing their individual assignments. Most observations were conducted in the student lab, in the presence of other students from the same and different classes. Observations followed the lines of contextual inquiry, in which the researcher follows the student's lead, asking clarifying questions and noting important occurrences [4]. Two students were interviewed after the SNA project using their diaries as the basis for the discussion. Observations and interviews were audio recorded and later transcribed.

Data Analysis

We began our analysis by compiling data into individual student profiles containing survey responses, diaries, observation notes and interview transcripts, assignments with peer and instructor comments, and grades. Based on initial observations, we created a list of open-ended questions that we intended to answer from the study data. These were elaborated on and modified as we began to create summary reports on selected issues and aggregate supporting data across students. These reports were the basis for in-depth analysis to identify common themes and patterns across students. We took the grounded theory approach, allowing the themes to emerge from the data [8], and narrowed our focus to two questions we deemed most pertinent to understanding the adoption of SNA by novice community analysts:

- *What process and phases emerged from the students' practices in analyzing the social media data?*
- *What factors affected the students' experience and ability to achieve their objectives in each phase of the process?*

We investigated these issues relative to the specific course assignments, teaching material, and teaching instructions but looked for evidence to support broader principles. We substantiate our findings by combining the quantitative data from the surveys, group modeling exercise, and diaries with qualitative data extracted from observations and interviews.

Study Participants

The study involved the instructor and 15 out of 19 students in the class. Based on the pre-study survey, the group of participating students (hereafter "students") included 2 male and 13 female students, 11 in the age range from 25 to 34, two younger than 25 and two older than 34. Most students are active social media users: 12 check a social networking site at least once a day and all but 2 participate in an online community at least once a week. Their experience with Excel varied. On a Likert scale with 1 indicating "complete novice" and 7 indicating "expert user," the median was 4, with the min=2 and max=6. Ten students used Excel at least once a week; 6 of whom used it daily. All but 2 students could add values in the spreadsheet and sort data, and most could use autofill and change number formats. Less than 3 could create macros, use an x-y plot chart, or create a pivot table. Only 1 student had studied or performed SNA before.

FINDINGS

In this section we present our findings by discussing (1) results of students' work, (2) a process model describing the emergent practices of students using SNA to analyze social media data, and (3) factors that supported and challenged students through each step in the process.

Student Products of Social Network Exploration

Inspection of the students' intermediary and final assignments showed that, with relatively minimal training and feedback, the students were able to create sophisticated and meaningful network visualizations that communicated important findings about their communities. This is not to say that the SNA project was easy. In-class and one-on-one discussions with the instructor and researchers made it clear that nearly all students thought this was one of the most conceptually and technically challenging assignments they had completed. Diaries revealed that students spent a median time of 12.5 hours outside of class on their SNA project, ranging from 5 to 25 hours. Much of this time was consumed by collecting data, learning SNA concepts, and learning how to use and troubleshoot NodeXL. Comparable work performed in the future would likely require less time.

Student efforts resulted in data analyses, visualizations, and insights that were as varied as the communities studied. Typically a student used one of only a few data structures and visualization "types" they learned in the class or from interaction with other students. However, they successfully modified these in non-trivial ways to match their community's data and context-specific goals.

The most common graph type was a bipartite graph, e.g., with nodes representing the community members and subgroups they belong to (see Figure 1). The next most common graph type was a directed graph representing community member discussions. Following the instructor's suggestion, two students created networks that show implicit relationships between entities, unlike most cases where connections were explicit in the data. Figure 2 illustrates this approach, by linking fashion designers to one another based on the number of people who listed both as favorites. Only 2 students created graph types that were not discussed in class, e.g., a bipartite graph of people and days.

While all students used standard data structures, their end results were dramatically different because many used unique, community-specific variables and attributes. "Subgroups" within different communities often meant very different things, ranging from discussion forums to wiki pages to craft swap groups. Visual properties of nodes, such as color and size, were used to represent important social roles, e.g., a network administrator or important community members, and corresponding SNA metrics, such as *in-degree* or *betweenness centrality* computed for a number of projects completed (Figure 1), or amount of lost weight in a weight loss discussion forum. Visual properties of edges, such as line thickness were used to indicate the strength of a tie between entities (Figure 1 and 2) or, through different colors, represent different types of connections. In all cases,

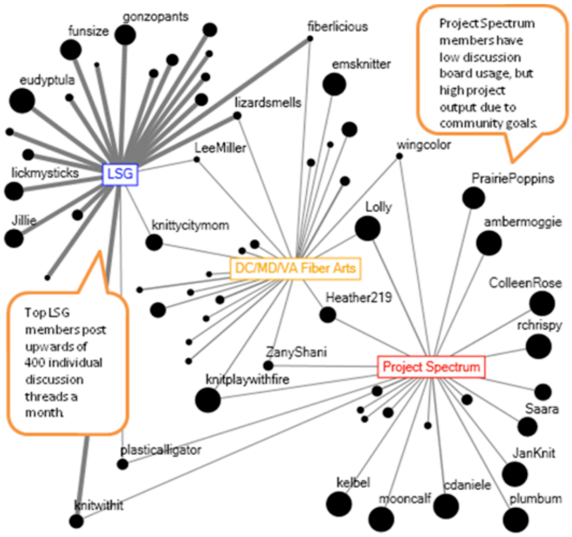

Figure 1: Subgroups in the "Ravelry" community. This bipartite graph shows community members (circles) connected to the subgroups (rectangles) they participate in. Node size reflects number of craft projects completed, while edge width indicates the number of posts to the subgroup forum.

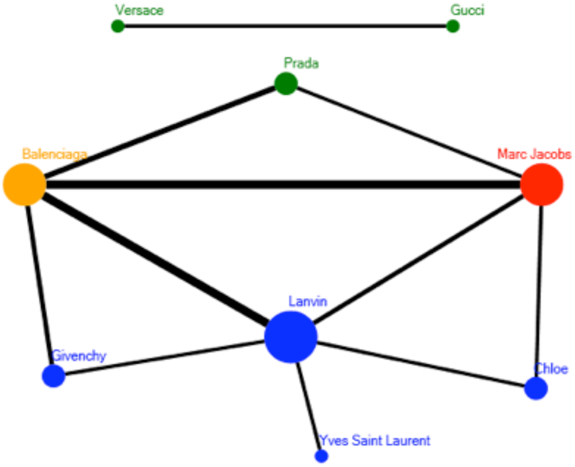

Figure 2: Fashion Designer Connections. This graph uses "favorite designer" data from 75 random Fashion Spot members to infer relationships between fashion designers. Node size indicates the degree of node centrality and edge width indicates the tie strength (i.e., the number of Fashion Spot members who marked both designers as a "favorite"). Edges with tie strengths under 5 were removed for clarity. Color indicates country of origin.

students mapped multiple variables of interest onto different visual properties in the same graph.

With a few exceptions, students' approached NetViz Nirvana and made important context-specific observations in their reports. The hard-earned but apparent success of the students suggests that SNA novices can learn and apply SNA effectively to expand their understanding of communities with moderate educational scaffolding.

Process Model of SNA & Visualization

As detailed earlier, 12 of the 15 students participated in a moderated discussion about the SNA process at the end of their projects. The model they collaboratively developed was elaborated on using diaries and interviews and dubbed NAV (see Figure 3). While NAV is only a *descriptive* model, the students' success suggests that it may be a good first approximation for a *prescriptive* model for SNA novices. Importantly, students developed NAV absent any knowledge of existing, generic sensemaking models.

Process Characterization

Students were strikingly similar in their characterization of the overall process. Ten of 12 students who completed the post-survey independently identified the Define Goals and the Learning SNA Tool phase; all 12 identified the Collect & Structure Data and the Interpret Data via Network Visualization steps, and 6 students identified the Interpret Data via SNA Metrics as a separate activity. Only a couple students explicitly identified the Prepare Report phase; however, it was referred to in many students' diaries. In fact, the diaries often mentioned NAV process phases, particularly Data Collection & Structuring, Interpreting via Visualizations, and Learning SNA Tool. Two individuals also suggested Getting Feedback from Others as a distinct activity. Upon deeper analyses, we saw that peer based feedback permeated the entire process (detailed later). Here we focus on two important characteristics of the observed process: the extensive iterations and refinements and the use of graph visualizations as the means of sensemaking.

Iterations and Refinements. The iterative nature of the model cannot be emphasized enough. During the collaborative class discussion of the NAV model, students emphasized that the work evolved organically: one activity led to another and sometimes resulted in a completely new analysis. In many instances, activities informed one another in a spiral of successive refinement. The frequent switching between activities was apparent in the diaries of students who kept detailed records, i.e., in 5 minute intervals.

For example, one student described refining her goals several times and collecting 4 rounds of data after visualizing the initial dataset and looking at other students' work. Other students didn't clearly define their goals until after they had viewed their data. A few students reformulated their hypotheses after initial data collection in ways that could not have been defined easily from the start. As one student explained, "*A really good research question isn't just stated once and carried through…you collect some data and you go back and reevaluate your question*."

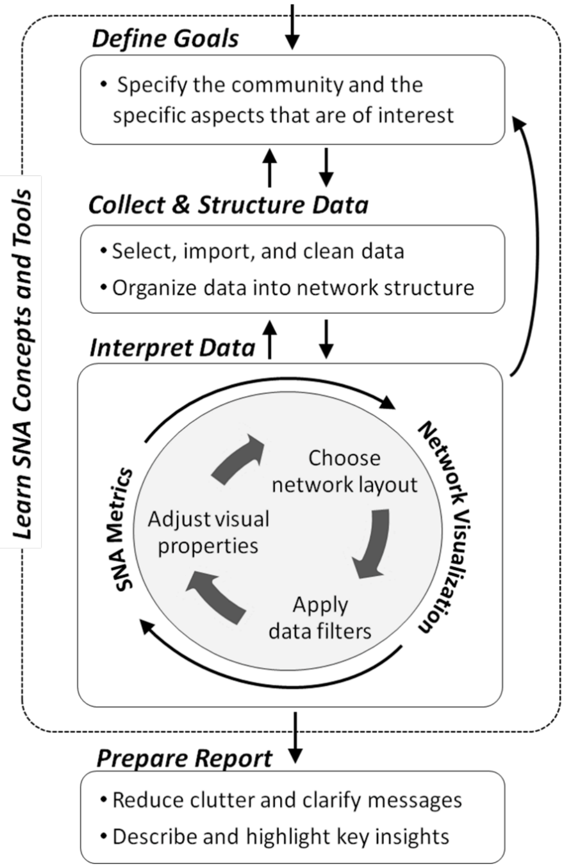
Define Goals
• Specify the community and the specific aspects that are of interest
Collect & Structure Data
• Select, import, and clean data
• Organize data into network structure
Interpret Data
Choose network layout
Apply data filters
Adjust visual properties
SNA Metrics
Network Visualization
Learn SNA Concepts and Tools
Prepare Report
• Reduce clutter and clarify messages
• Describe and highlight key insights

Figure 3. Network Analysis and Visualization (NAV) Process Model: steps and activities derived from the students' practices in analyzing community data using SNA metrics and NodeXL tool.

Impact of Visualization. The role of visualizations was essential throughout the process. In the exploratory cases outlined above, visualization provided key insights that helped the students assess whether they were on the right track with data collection and goals, e.g., it "*helped put order to a chaotic place like a message board*." However, recognizing that order often began with an initial visualization that was "*kind of a mess" and "very difficult to read*." The process of bringing order to the data was highly visual, as one student emphasized: "*Seeing the Senators' clusters during the tutorial…was a watershed moment…So I just started playing with the other data. It was fun. It was Social Network Illustration*."

Factors Affecting Student Experience

Students faced many challenges and worked around them as they completed their assignments. Several factors affected their experience, and collaboration with peers and the instructor played a critical role. We discuss them for each high-level phases of the NAV model (Figure 3).

Define Goals	
CH:	*Overwhelming number of choices; Generating a "network question"*
S:	*Instructor guidance and seeing peers' visualizations*
Collect and Structure Data	
CH:	*Onerous task of manually collecting data; Understanding how to structure network data*
S:	*Coupling of data and visualization in NodeXL helps with preliminary investigation of data structure; Peer feedback and knowledge exchange.*
Interpret Data using SNA Metrics	
CH:	*Lack of insights and working experience with SNA metrics; fear of complex calculations.*
S:	*Coupling of data and visualization in NodeXL supports exploration and gaining insights about specific metrics; Instructor guidance and feedback.*
Interpret Data through Network Visualization	
CH:	*Overwhelming number of possibilities for mapping data parameters into visual properties of the graph; Lack of experience and established practices.*
S:	*Peer feedback and knowledge exchange.*
Prepare Report	
CH:	*Manual modification of networks to optimize layout can be onerous for complex graphs.*
S:	*Student and instructor feedback available but not solicited during this stage.*
Learn SNA Concepts and Tools	
CH:	*Transitioning from simplistic examples in tutorial to large, messy datasets; Excel proficiency; Troubleshooting the NodeXL software; uncertainty about the cause of the problems.*
S:	*Easy explorations of data through NodeXL. Peer support and instructor guidance.*

Table 1. Challenges (CH) and Supports (S) for each phase in the process model

Define Goals

Most students (8/12) identified Defining Goals as the hardest conceptual step in the process. When asked what advice they would give to others working on the same SNA assignment, 7 of 12 students recommended defining specific goals and outlining the analyses before collecting data. The seemingly endless possibilities for analysis, given the open-ended nature of the assignments, were sometimes overwhelming. From an educator's perspective, however, minimizing this struggle may actually deprive new users from a necessary and effective stage in the learning process. Classroom discussions, draft ideas posted to the website, and instructor's feedback, revealed that initially, some students did not know what types of questions were amenable to SNA. For example, several students developed questions about correlations: *Do people who post more often also complete more projects?* It took some time and the instructor's help to arrive at a formulation of the question that connects with the network structure and metrics in a meaningful way (e.g., as in Figure 1). Interviews and post-surveys suggest that seeing other students' visualizations "*opened up a realm of possibilities*" for some and was often enough to help someone devise a "*network question*" for their community. Thus, both expert and peer advice was useful at this stage.

Data Collection & Structuring

The major sub-tasks for this activity include: (1) browsing through the site to determine what data could be reasonably collected, (2) deciding what to collect (closely related to Define Goals phase), (3) collecting data from the community, which occasionally included hand-coding messages into categories of interest, (4) restructuring the data into an edge list, and (5) adding attribute data into appropriate NodeXL worksheets.

Collecting data was primarily a manual process of hand-entering information found on member profiles or other community pages such as discussion forums. This manual process took on average 2.5 hours as recorded in student diaries. Data collection was described as "tedious" by most, although some were initially excited in anticipation of subsequent analyses. In class, several students asked to learn Excel shortcuts and formulas to help collect and format data. One student sought help from a colleague in Computer Science who wrote scripts for data cleaning.

A conceptually challenging aspect of this phase was the understanding how to structure network data to realize a desired graph. The students explored various options in order to expose the types of relationships they wanted to study (e.g., Figure 2). When discussing how to structure the data they often described the desired visualization. In that respect, the close coupling of data and visualization in NodeXL was invaluable. It enabled them experiment with graph construction, manipulate data via visual properties, and develop insights about the relations between the two.

On a practical level, some data networks were very challenging to represent for non-programmers. For example, the student who studied the network in Figure 2 created the edge list and added tie strength values manually. That approach does not scale and very quickly increase in complexity with the number of nodes.

In addition to the complexity of data structuring, the quiz revealed students' confusion over the difference between the data stored in the Edges worksheet versus the data in the Vertices worksheet. Four students needed assistance from the instructor to get their data into an appropriate edge list.

Overall, students required additional lab support on ways to structure the data and how to organize it within NodeXL.

Interpreting Data via SNA Metrics

The major sub-tasks associated with this phase included (1) calculation of network metrics within NodeXL, (2) sorting vertices based on the metrics, i.e., to identify individuals with the highest eigenvector centrality, and (3) mapping metrics to visual properties of the graph.

Students' experience with network metrics was mixed. In the final report, only 1 student visually represented metrics other than *degree* and 2 others only mentioned them in the write-ups, despite the encouragement from the instructor to include the analysis. This is partially due to the fact that many students used bipartite graphs for which many of the metrics, e.g., *betweeness centrality*, do not make sense. However, two students mentioned that they didn't focus on metrics because they "*didn't have time to think about the more complicated metrics*." Another student described how "*seeing all those numbers*" dredged up memories of math anxiety. Thus, in this phase, the peer help was not as useful as the expert's help, because of students' unfamiliarity with SNA metrics and the lack of self-confidence that they can apply them correctly.

Interpreting Data via Visualizations

The major tasks associated with this stage include: (1) choosing an initial layout and graph type, (2) setting visual properties, e.g., the edge width, vertex size, color, and shape, to express various data properties, (3) filtering edges and vertices, (4) displaying labels, (5) calculating and viewing clusters, and (6) comparing multiple visualizations.

The role of visualization in interpreting data was crucial. Most students (9/12) mentioned that visualizations changed their understanding of the community either somewhat (3) or significantly (6). Even when asked about the general role of SNA, not just focusing on visualizations, half of the students mentioned the benefit of "seeing" relationships that might not have been apparent otherwise, e.g., the importance of boundary spanners that connect two clusters of nodes. Thus, for student as novice analysts, the visualizations were far more important than metrics.

Comments on draft visualizations, received from peers and the instructor, were reported to be helpful. Indeed, most students (7/12) thought that, in this stage in the process, the feedback from peers was most helpful, and 4 thought the help from the instructor was most helpful. We note that the peer comments on drafts focused primarily on improving visual properties and layouts, sometimes recommending new variables to map onto visual properties.

Creating or analyzing visualizations was identified as the "most rewarding" activity by 11 out of 12 students. Some students liked analyzing and interpreting the visualization, while others found satisfaction in reaching NetViz Nirvana. Students often used words like "*pretty*" and "*beautiful*" to describe their visualizations. The aesthetic nature of this work, mixed with the creative process, seems to have resonated well with the students.

NodeXL's ability to manipulate nearly every visual aspect of the network was a bit of a two-edged sword. At first, students felt overwhelmed by the possibilities. They struggled to know which visual properties to assign to which variables, sometimes overcomplicating the graph by using several visual properties for a single attribute. However, during the course module, some best practices started to emerge, such as using different shapes to indicate different types of nodes in bipartite graphs instead of simply using different colors. Although students did not always know at first how to best represent a particular attribute, they often recognized it when they saw it.

Prepare Report

Final student reports included at least 2 visualizations and text describing their importance and meaning. The preparation of the report had two key phases: (1) fine-tuning the visualizations to meet the NetViz Nirvana visual principles as close as possible, and (2) describing the final visualizations and insights learned.

Manual fine-tuning of visualizations took a considerable amount of time. Most students (7/12) said that creating visualizations and, in particular, adjusting layouts was one of the most technically challenging activities. But, it was not considered conceptually challenging. According to the diary entries, most students spent over an hour adjusting layouts. A typical approach was to use a built-in layout algorithm, e.g., Fruchterman-Reingold, and then adjust the individual nodes manually.

Overall, students were quite effective in interpreting their visualizations. They often relied upon a broader knowledge about the observed communities to explain why the visualization made sense. Only one student included annotations on the final graph (Figure 1), despite the instructor's emphasis that this is potentially a useful strategy. This may be due to the fact that NodeXL did not directly support annotations of the graph images. One could add them only when writing a report, using document editing facilities. Likewise, NodeXL did not include a legend to link visual properties, such as color, to concepts or titles. A few students ran into another technical challenge: the lack of support for multiple graphs within the same NodeXL file. This was frustrating for a handful of students who created visualizations that relate to one another in some way. They had to create two files and fine-tune the visual properties of each file independently.

Students were generally satisfied with their final reports, as evidenced by their willingness to share them widely through public posting and diary entries. Students did not solicit input from peers or the instructor during this stage.

Learn and Troubleshoot the SNA Tool

Most students (8/12) mentioned that the use of the NodeXL technology left them confused and uncertain about their work. The usability aspects of the tool are described elsewhere [5]. Here we focus on the learning objectives and the troubleshooting of NodeXL as part of the entire process.

Many students expressed sentiments similar to those of a student who said that NodeXL "*is very strong, very nuanced, but not very approachable*." She said she felt as if she had completed one cooking class and was given a set of very sharp knives to use. This is likely because the students did not have sufficient time to develop a clear mental model of network analysis. This led to confusions with the technology, from the basic understanding where the data should be in the spreadsheet, to interpreting concepts like "tie strength". Having many different ways to accomplish the same task also confused students at times.

Many students (7/12) said that getting help from the instructor was most helpful when learning Excel skills and NodeXL features. Novice Excel users had trouble using formulas and entering data, and were not comfortable using time-saving shortcuts, like autofill. Only 2 students acknowledged that peers have been helpful in learning the technology but observations of the lab session made it clear that students helped each other more often than they remembered. The resource they most used throughout the module was the NodeXL Tool tutorial. The tutorial was helpful to students, although some complained that the datasets used in examples were too simple and insufficient. Thus, the students struggled to which made translating the ideas to their communities challenging.

Another challenge that students faced were error messages and sporadic software failures. For the study we used an earlier version of NodeXL and thus most students ran into at least one error message or crash. This dampened their excitement for the assignment. It was particularly hard because when students didn't get the desired result they wondered if it was because of the software or because they had done something wrong: "*So then I start doubting myself*", a student said.

DISCUSSION

Process Model Comparison and Discussion

Our NAV process model (Figure 3) extends the sensemaking models that were derived by analyzing experts in different contexts [16, 19, 22]. Similarity between models exists in the iterative nature of data collection and analysis and a gradual progression towards an increasingly insightful synthesis of data findings. For example, the popular sense making loop NVAC (Fig 2.1, page 43, 22) refers to four iterative stages: Gather Information, Re-represent, Develop Insights, and Produce Results. Likewise, our NAV model includes the analogous phases: Define Goals, Collect & Structure Data, Interpret Data, and Prepare Report (Figure 3). These parallels suggest that SNA is primarily a sensemaking activity and that we may derive inspiration from the broader sense making literature.

However, the experiences of our student analysts highlighted the importance of explicitly considering the learning process throughout the model as captured in our phase for Learning SNA Concepts and Tools. Indeed, the learning activities in NAV may complement, but are distinct from the Interpret Data phase or Develop Insights in NVAC. Generally, the learning phase for novices would correspond to the need of experts to identify new conceptual frameworks and master new tools in order to enact the rest of the sensemaking process. We postulate that this is harder to detect with experts unless the study is conducted over a longer period of time or in non-routine situations where existing expertise is not sufficient. Thus, our novice users revealed an essential, often overlooked aspect of the sensemaking process.

As for mastering exploration techniques in the SNA context, the NAV model highlights the importance of *visualization* as a mechanism that enables users to conceptually connect the data, SNA techniques, and tool affordances. It served as a foundation for building a common vocabulary. Students rarely invoked network concepts without drawing pictures or using visual language. Simple data plots provided some insights, but the true conceptual value of integrated visualizations was shown when students tied visual properties to metrics and data cells. And seeing others' visualizations helped students recognize what was possible and learn how to structure data and ask questions amenable to network analysis. Like Klein et al.'s [16] *frames*, network visualizations helped manage attention, define, connect, and filter raw data. However, unlike *frames*, network visualizations were not triggered by unexplained phenomena and their plausibility was not questioned [16]. Instead, their usefulness and insightfulness were continually scrutinized, both through the learning process and assessment of results.

We anticipated that challenges would invoke strong emotional reactions with students. Similarly to the Kulthau's work on information seeking [15], we considered affective aspects of the SNA experience. The strong feelings identified from user diaries often reflect students' struggle with an unfamiliar task and sensemaking process. Feelings of uncertainty, for example, accompanied unclear and underspecified project goals while excitement and increased interest were associated with the initial viewing of data visualizations and completion of a task. The affective dimension of our investigation enriched our analysis and suggests that we can reduce the anxiety and sense of being overwhelmed by improving reliability, providing an 'undo' feature, and providing better layouts and default settings to reduce manual adjustments.

Implications for Designers

Tools like NodeXL have removed the complexity of programming, previously required to analyze and visualize network data. Rich SNA tasks can now be achieved simply by a selection of NodeXL functions and quickly leads to a rewarding experience of visual exploration and iterative refinement. The NAV process model (Figure 3, Table 1) reveals the importance of visualization and analysis support provided in the sense making process. However, it also suggest that SNA tools could become more effective if designed to encompass and seamlessly integrate more steps of the NAV process. Here we reflect on these opportunities.

Improving Data Collection. Similarly to other tools, data collection in NodeXL is based on "*import first, then analyze*" model, which assumes that the analyst knows what data to collect right from the start. Our study shows that is not always the case, particularly not with novice analysts. Thus, it is important to find a way to connect the tools with the data sources in a way to facilitate dynamic expansion and filtering of datasets with the aid of facilities for visualization and metrics computation. This would allow users to develop goals for analysis iteratively and gain insights into the nature of the data without iterative import functions. For example, a general-purpose network browser could be developed (similar to [14]) to facilitate customized data capture and updating. The browser could provide summary graphs for specified attributes [23] to help analysts determine what data to analyze in-depth.

Improving Report Preparation. Network visualizations can serve as an effective means of communication, particularly when easily readable and aesthetically pleasing. Students found the task of preparing the final reports highly rewarding. However, the amount of time spent to refine the layout and approach the NetViz Nirvana criteria was seen as excessive. Observations indicated that default parameters for the layout are critical to avoid tedious manual interventions. Also, providing more varied layouts with clear visual quality measures, like those in Social Action [18], would encourage the creation of more effective graphs by novices. Significant improvements may benefit from alternate input devices that let users more naturally reposition nodes and node clusters. Alternatively, human manipulation of node placement could serve as the basis for a machine learning layout algorithm.

Supporting collaboration. Currently, most of the SNA tools are stand-alone programs with a very few built-in collaborative features. Our study shows that both the peer and expert feedback are instrumental in acquiring new skills and learning how to apply them effectively. These findings provide strong empirical support for the design of services such as ManyEyes [25] which promote sharing of best practices and community feedback on submitted visualizations. Our NAV process model suggest that such communities may want to explicitly encourage discussions of specific NAV stages such as SNA goal setting, structuring data into a network that supports the target analysis, the interpretation of graphs and metrics, etc. They may also want to organize the submitted visualizations based on (1) commonly recognized graph characteristics, e.g., bipartite graphs, directed graphs, etc., (2) data types, such as wiki, forum, protein and other biological structures, or (3) quality measures such as NetViz Nirvana quality metrics [10]. Furthermore, the users willingness to undertake manual graph refinements suggest that large scale graphs could be manually fine-tuned through a crowdsourcing tool that allows a community of practitioners to "clean up" different sections of a network, particularly if the tedious task is made into an engaging social game or supported by similar incentives.

Implications for Educators

Structuring SNA Assignment. The SNA module assignment achieved the objective of developing student competencies in all SNA stages and becoming proficient in using a tool that can aid that process. Indeed, students specified their own research goals through a conceptually challenging, yet a very effective learning exercise. They had to figure out on their own what data to collect and how to structure it for effective network analysis. Gaining experience and confidence in that aspect of SNA was one of the most important outcomes. In contrast, time consuming manual collection and structuring of data did not have any direct payoffs. For non-technical users, the instructors may want to (a) provide datasets that are prepared to allow for creativity and flexibility, at the expense of students' freedom to work with a community of their choice, (b) provide automatic import tools from online communities, e.g., Twitter and, Facebook, or (c) instruct students how to use basic screen scraping tools and scripts. It may be advisable to begin with a primer on Excel features and exercises in mapping data into networks. Generally, we recommend the collaborative classroom set up. The study showed that the exchange of draft visualizations and peer feedback in an online forum was very helpful and confirmed that novices can meaningfully aid other novices and thus collectively accumulate valuable know how.

Introducing Network Concepts. The close coupling of the spreadsheet data and visualization in NodeXL was helpful when introducing basic network concepts. The instructor was able to effectively describe network concepts by displaying corresponding metrics and visualizations side-by-side on a projection screen, as well as highlight nodes of interest by selecting them. Mapping network metrics (e.g., degree) onto visual properties (e.g., size) of the simple kite network was an effective way to visualize each vertex's metric at the same time. The importance of the visualizations in understanding centrality measures and other network metrics cannot be overstated for novices.

Implications for Online Community Analysts

Existing tools, like NodeXL, make SNA increasingly accessible. Our study showed that with a moderate amount of time and educational scaffolding, students without strong quantitative background are able to apply SNA effectively and gain further insights into observed communities, e.g., identify unique and important individuals, subgroups, and overall community structure that were not otherwise apparent. However, the process of learning to apply SNA methods was hardly trivial. New analysts should expect a conceptually challenging experience, particularly when formulating "network questions" and structuring the data. The uncertainties at the beginning of analysis, tedious fine-tuning of visualization, and manual collection of data can be overwhelming. However, analysts can expect exciting moments of discovery and rewarding end results. And remember, the process may take several iterations.

CONCLUSION

We have reported on an in-depth case study of 15 graduate students who used NodeXL to learn SNA concepts and apply them to study online interaction. Their successes demonstrate that novice analysts can effectively adopt and apply SNA techniques within a relatively short time. We characterized their practices by the Network Analysis and Visualization (NAV) process model that enables us to articulate challenges in sensemaking of network datasets. Our study can serve as an empirical foundation for new designs of SNA tools and educational practices intended to support novices as well as a broader community. We hope it will provoke further discussion about SNA-specific sensemaking models and inspire new ways to support social media analysts. We may not make SNA as easy as Amazon checkout, but we believe that teaching the NAV process model and designing tools and services that embody that models will greatly facilitate analysts' work.

REFERENCES

1. Adar, E. GUESS: a language and interface for graph exploration. In *Proc. CHI 2006*, ACM Press (2006).
2. Adar, E. and Adamic, L. Tracking information epidemics in blogspace. *Proc. Int. Conference on Web Intelligence 2005,* ACM Press (2005), 207-214.
3. Barabasi, A. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, NY (2003).
4. Beyer, H., and Holtzblatt, K. *Contextual Design: A Customer-Centered Approach to Systems Designs*. Morgan Kaufmann, San Francisco, CA, USA (1997).
5. Bonsignore, E.M., Dunne, C. Rotman, D., Smith, M., Capone, T. Hansen, D. and Shneiderman, B. First Steps to NetViz Nirvana: Evaluating Social Network Analysis with NodeXL. *IEEE International Symposium on Social Intelligence and Networking (SIN-09)* (2009).
6. Burt, R. S. *Brokerage and Closure: An Introduction to Social Capital*. Oxford Univ. Press, New York (2007).
7. Carter, S., and Mankoff, J. When participants do the capturing: the role of media in diary studies. *Proc. CHI 2005*, ACM Press (2005), 899-908.
8. Corbin, J. and Strauss, A.C. *Basics of Qualitative Research:Techniques and Procedures for Developing Grounded Theory.* Sage, Thousand Oaks, CA (2007).
9. Cross, R., Parker, A. and Cross, R. *The Hidden Power of Social Networks*. Harvard Business School Press, Cambridge, MA (2004).
10. Dunne C., and Shneiderman, B. Improving graph drawing readability by incorporating readability metrics: A software tool for network analysts. HCIL Tech Report HCIL-2009-13. University of Maryland (2009).
11. Freeman, L. Visualizing Social Networks. *J. of Social Structure*, 1 (2000).
12. Google Analytics: http://www.google.com/analytics.
13. Hansen, D., Shneiderman, B., & Smith, M. Network Analysis with NodeXL: Learning by Doing. http://casci.umd.edu/NodeXL_Teaching
14. Heer, J. & Boyd, D. Vizster: Visualizing online social networks. *Proc. of IEEE Info. Visualization*. (2005).
15. Kuhlthau, C. C. *Seeking meaning: A process approach to library and information services*. Libraries Unlimited, Westport, CT (2004).
16. Klein, G., Phillips, J. K., Rall, E. L. and Peluso, D. A. A Data-Frame Theory of Sensemaking. *Proc. of the Sixth International Conference on Naturalistic Decision Making*. CRC Press (2007), 113-155.
17. Lithium Technologies. Community Health Index for Online Communities. White Paper (2009).
18. Perer, A., & Shneiderman, B. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. *Proc. IUI '08*, ACM (2008), 109-118.
19. Pirolli P., & Card, S. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Through Cognitive Task Analysis. Paper presented at *Int. Conf. on Intelligence Analysis* (2005).
20. Smith, M., Shneiderman, B., Milic-Frayling, N., Rodrigues, E. M., Barash, V., Dunne, C., et al. Analyzing Social (Media) Network Data with NodeXL. *Proc. Communities & Technologies* (2009).
21. Telligent Analytics^{TM} reporting server: http://telligent.com/products/telligent-analytics/
22. Thomas, J., & Cook, K. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, Los Alamitos, CA (2005).
23. Wasserman, S. & Faust, K. *Social Network Analysis*. Cambridge University Press (1994).
24. Wattenberg, M. Visual exploration of multivariate graphs. *Proc. CHI 2006*, ACM Press (2006).
25. Wattenberg, M., Kriss, J., and McKeon, M. ManyEyes: a Site for Visualization at Internet Scale. *IEEE Trans. on Vis. and Computer Graphics 13*, 6 (2007), 1121-1128.
26. Welser, H., Gleave, H., Fisher, D., and Smith, M. Visualizing the signatures of social roles in online discussion groups. *J. of Social Structure, 8*, 2 (2007)