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

Derek L. Hansen*, Dana Rotman*, Elizabeth Bonsignore*, Natasa Milic-Frayling^, Eduarda Mendes Rodrigues^, Marc Smith+, Ben Shneiderman*

*University of Maryland, Human Computer Interaction Lab; ^Microsoft Research; +Telligent Systems

ABSTRACT

Voluminous online activity data from users of social media can shed light on individual behavior, social relationships, and community efficacy. However, tools and processes to analyze this data are just beginning to evolve. We studied 15 graduate students who were taught to use NodeXL to analyze social media data sets. Based on these observations, we present a process model of social network analysis (SNA) and visualization, then use it to identify stages where intervention from peers, experts, and computational aids are most useful. We offer implications for designers of SNA tools, educators, and community & organizational analysts.

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 tools and services have enabled new forms of collaboration and activities in nearly every human endeavor imaginable. Companies reach passionate consumers, scientists share data, genealogists find common ancestors, and online community members discuss topics ranging from entertainment to politics to world hunger. We expect more novel and rich technology-mediated social interaction as social media reaches new populations and goes mobile.

Although we often emphasize social media success stories, we must remember the countless failures due to both social and technical factors. How can we support practitioners in their endeavor to cultivate sustainable online communities? 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 their website traffic so they can make more informed decisions. We envision an equivalent set of *social analytics tools* (e.g., [16, 21]) to help online community analysts make better decisions based on a deep understanding of social participation and relationships.

Social network analysis (SNA) provides a set of concepts and techniques for making sense of social data through quantifiable metrics and network visualizations. These complement basic metrics of social participations used in current tools (e.g., number of posts; membership duration) [16, 21] and reveal the patterns in the network that result from social interactions. SNA concepts provide an effective vocabulary to characterize important relational properties of network members, as well as entire network structures. However, SNA also adds significant complexity and imposes obstacles for analysts that lack technical skills.

Our long-term goal is to develop simple, yet powerful social analytics tools and educational strategies for online community analysts, participants, and designers. In this paper, we make two primary contributions toward that goal. First, we present a process model describing how a group of student analysts constructed insights into social media datasets through the use of SNA metrics and visualizations. Second, we use this model to identify stages in the process where intervention from peers, experts, and computational aids are most useful. We then offer design ideas and teaching strategies for making SNA tools and services more collaborative and accessible to novice and expert users.

RELATED LITERATURE

SNA and its mathematical companion graph theory have a long and distinguished history in the academic world [2, 11]. In recent years, many researchers have used SNA to examine social interaction in computer-mediated environments, helping to identify unique social roles [24], social structures [6], and information dissemination patterns [1]. Despite SNA's success in academic circles and its appearance in mainstream publications (e.g., [2]) and management literature [9], it has not been widely used by practitioners. This is likely due to the lack of usable SNA software, lack of education, and historical challenges of collecting social network data. As more usable SNA tools are developed and taught, and data from social media become ubiquitous, SNA can reach a much wider audience.

To fully realize the potential of SNA, it is important to understand the process novices go through to make sense of social data from virtual organizations and communities. Process models that describe key activities, tasks, cognitions, and/or feelings have been useful in helping design novel tools [18] and educational interventions [14].

Process models are particularly good at identifying moments where interventions from peers, experts, or computational aids are most useful. Pirolli and Card call these moments "leverage points" [18] in their process model of information analyst's activities. Their work is part of a larger endeavor to characterize the sensemaking process of intelligence analysts [15, 22]. In a different, but related domain, Kulthau maps a process model of information seeking behavior onto Vygotsky's concept of a "Zone of Proximal Development" to identify stages in the process at which information mediators (i.e., educators) can help students the most [14]. We take our inspiration from these and related process models. Like Kulthau's original studies, we focus on novices, however we are interested in the entire sensemaking process rather than information seeking per se. Our model is closest to sensemaking models [15,18,22], although existing sensemaking models are based on experts and have yet to explicitly characterize the role of visualization as has been recently suggested [22].

The development of SocialAction [17] 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. betweenness, plotted on a scattergram), (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 serve not only to frame the process experts follow when exploring complex data sets, but also to guide students as they engage in SNA about their own communities of interest.

Most work on visual analytics has focused on individuals. However, systems such as ManyEyes have begun to explore ways of supporting distributed, collaborative sensemaking of data visualizations [23]. We believe that identifying the role of collaboration throughout the sensemaking process can lead to novel design ideas for collaborative SNA and visualization tools.

METHODS

In order to observe the adoption of SNA metrics, tools, and practices, we conducted an in-depth month-long study of 15 students in a graduate course on Computer-mediated Communities of Practice (CoP).

Study Setting

The study leverages the availability of NodeXL, a free, open-source SNA tool (http://www.codeplex.com/nodexl) and the University of ________'s teaching environment.

NodeXL SNA Tool

The NodeXL add-in for Excel 2007 is an open source extension to the widely used spreadsheet application that provides a range of basic network analysis and visualization features [20]. NodeXL uses a highly structured workbook template that includes multiple worksheets to store all the information needed to represent a network graph. Network relationships (i.e., graph edges) are represented as an "edge list", which contains all pairs of entities that are connected in the network. Complementary worksheets contain information about each vertex and cluster. Visualization features allow users to display a range of network graph representations and map data attributes to visual properties including shape, color, size, transparency, and location. Although NodeXL is in a perpetual beta release, it was selected for teaching SNA because of its perceived ease of use, familiar spreadsheet paradigm, coverage of core SNA metrics, and rich support of network visualizations.

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 hybrid course includes a weekly in-person 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 teaching module that introduced SNA concepts, NodeXL, and their application to the data from online communities through in-class and online discussions and independent readings. The module occurred 1/3rd of the way through the CoP course and required them to analyze a community they had been following as participant observers 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 was based on a process model with 9 steps: (1) basic, (2) layout, (3) visual design, (4) labeling, (5) filtering, (6) grouping, (7) graph metrics, (8) clustering, and (9) advanced.

Course requirements included a take-home quiz and an intermediate and final report in which students used NodeXL to analyze data from their chosen community. Course readings and discussions of SNA concepts covered community metrics, social roles, and network visualization quality as measured by NetViz Nirvana [10], which states the following recommendations: (1) Every node is visible, (2) for every node you can count its degree, (3) for every edge you can follow it from source to destination, and (4) clusters and outliers are identifiable.

Students were encouraged, but not required, to participate in the study. Precautions were made to assure that participation would not impact students' grades. Participants received a hardware or software gift.

Data Collection

In order to capture a wide range of aspects about the SNA teaching and learning, we collected and analyzed data from a variety of sources as outlined below.

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.

Student Assignments. Researchers analyzed all SNA assignments, instructor comments and grades, and peer and instructor comments on the course website. Assignments included (a) a take-home quiz that required students to use NodeXL to answer questions about a user interaction graph derived from an internal discussion forum of a company, (b) intermediate SNA visualization(s) 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 (using the "NetViz Nirvana" as a guide [10]), the accuracy and quality of the description, and the importance of the analysis in understanding the community. Students were encouraged to help one another and two voluntary lab sessions were set up to facilitate peer support.

Surveys & Group Modeling. Prior to the start of the SNA module, the course instructor assessed each student's level of Excel, SNA, and social media literacy with an online survey. A hand-written, post-survey was administered during class on the day students turned in their final SNA assignment. One of the research members also oversaw a collaborative process mapping 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 unified class-wide process model that was drawn on the chalkboard and discussed. Students used the model to answer questions identifying the most challenging tasks, as well as tasks where peer and expert advice was most helpful.

Diaries. While working on the SNA quiz and assignment, students completed 3 semi-structured diaries. Self-reporting diaries have been used to help inform process models before [14], and are gaining popularity in technology studies, allowing students to capture their actions and emotions with little interference from the researcher [5, 7]. Our diary form included open-ended questions about the work that was performed, the role of peer and expert support, resources that were used, and their satisfaction with their own work. Diaries also provided a space for students to detail the tasks they performed, the amount of time spent on the tasks, and their feelings during those tasks. Many students were observed filling out the diaries as they worked on their assignments in the lab, although some appear to have filled them out later.

Individual Observations and Interviews. One of two researchers observed or interviewed each student. Eleven students were observed while completing their individual assignments. Most observations were conducted in the student lab, while other students (from the same and different classes) were also present. Observations followed the lines of contextual inquiry, in which the researcher follows the student's lead while asking clarifying questions, and noting important occurrences [3]. Two students were interviewed after the SNA project. Their diaries were used as the basis for the interviews. Most observations and interviews were audio recorded and later transcribed. Detailed observer notes were used in cases where audio recordings were not available.

Data Analysis

We began our analysis by compiling data into individual student profiles containing survey responses, diaries, observation notes and interview transcriptions, assignments with peer and instructor comments, and grades. Based on the initial observations we created a list of open-ended questions that we were hoping to answer from the study data. These were elaborated on and modified as we began to create summary reports for selected issues by aggregating supporting data across students. These reports were the basis for in-depth analysis and group discussions in order to identify common themes and patterns across students. We took the grounded theory approach, allowing the themes to emerge from the data [8].

We ultimately focused on two questions most pertinent to our objective of understanding the issues in broad adoption of SNA and successful application to social media data:

What process and phases transpired 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 investigate these issues relative to the specific course assignments, teaching material, and teaching instructions but look 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 of the 19 students in the class. Here we describe the participating students (hereafter "students") based on pre-survey results. Two students were male, 11 are in the age range of 25-34, 2 are younger and 2 are older. Most students are active social media users (e.g., 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). Experience with Excel varied. On a Likert scale with 1 indicating "complete novice" and 7 indicating "expert user," the min = 2, max = 6, and median = 4. Ten students used excel at least once a week, 6 of who used it daily. All but 2 students could add values and sort data, most could use autofill and change numbers to a date/time format, and less than 3 could create macros, use an xy plot chart, or create a pivot table. Only one student had studied SNA as part of a course and none had performed SNA before. In summary, students were SNA novices with a range of Excel skills and significant experience with social media.

FINDINGS

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

Student Products of Social Network Exploration

Inspection of the intermediary and the final students' assignments show 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 projects 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" learned from class or from other students' preliminary visualizations. However, they successfully modified these it in non-trivial ways to match their community data and context-specific goals.

The most common graph type was a bimodal graph, e.g., with nodes representing the community members and subgroups they belong to, like in Figure 1. The next most common graph type was a directed graph representing community member discussions like the one used in the quiz and one of our readings. 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. The graph in Figure 2 illustrates this approach, by linking fashion designers to one another based on the number of people who listed both as favorites. Although the idea for the graph came from the instructor and a related graph structure in the tutorial, the student implemented it correctly in a meaningful and insightful way. Only two study students experimented with graph types that were not discussed in the class (e.g., a bimodal graph of people and days).

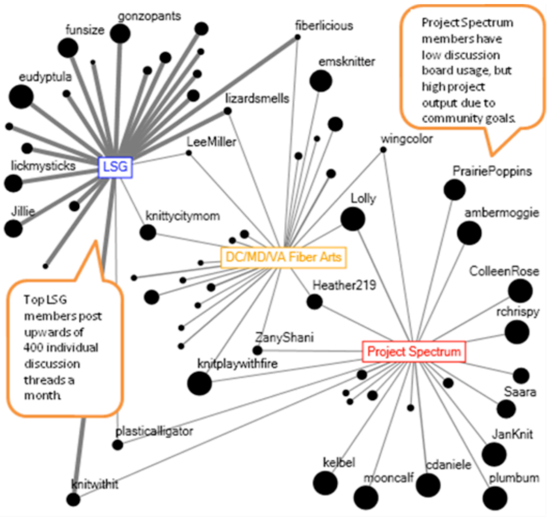

Figure 1: Subgroups in the "Ravelry" community. This is a bimodal network graph showing community members (circles) connected to subgroups (rectangles) they participate in. Node size indicates number of craft projects completed, while edge width indicates the number of posts to the subgroup forum.

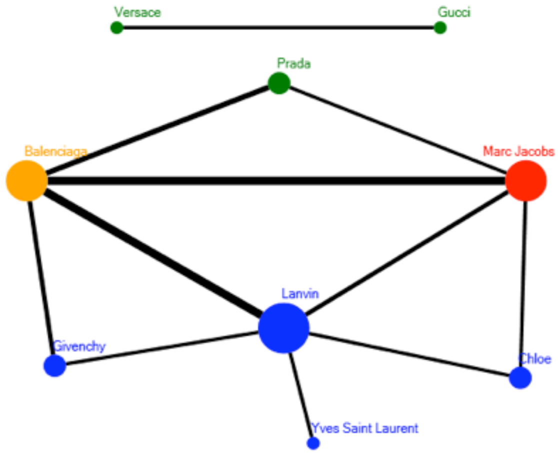

Figure 2: Fashion Designer Connections. This graph uses "favorite designer" data from 75 random Fashion Spot members to infer relationships between fashion designers. Node size indicates the degree of node centrality and edge width indicates the tie strength (i.e., the number of Fashion Spot members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

Although students used standard data structures, the end results were dramatically different because they used so many 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 attributes of nodes such as color and size were used to represent important social roles (e.g., administrator), SNA metrics such as betweeness centrality or degree, number of projects completed (Figure 1), and amount of weight lost (in a weight loss discussion forum). Visual attributes of edges were often used to indicate the strength of a tie in some manner as Figure 1 and 2 show, and some people used colored edges to represent different types of connections. In every case, multiple variables of interest were mapped onto various visual attributes in the same graph.

With a few exceptions, students' visualizations approached NetViz Nirvana and their descriptions identified important context-specific observations. Overall, the hard-earned but apparent success of the CoP 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 under "Data Collection," one of our final activities was a moderated discussion with 12 out of 15 students in which they arrived at consensus on a group SNA process model. We later elaborated on the model after examining diaries and interviews. Figure 3 shows the results. While this is only a *descriptive* model, the success of the students suggests that it may be a good first approximation for a *prescriptive* model for SNA novices.

Process Characterization

Students were strikingly similar in their characterization of the overall process. Ten of the 12 students who completed the post-survey independently identified the Define Goals concept and the Learning SNA Tool concept; all 12 identified the Collect & Structure Data concept, as well as the Interpret Data via Network Visualization concept; and 6 of the 10 identified the Interpret Data via SNA Metrics concept. Only a couple students explicitly identified the Prepare Report concept, however, in their diaries many more students identified that activity. Diaries also supported the other categories, with a particularly clear emphasis on Data Collection & Structuring, Learning SNA Tool (including discussions of troubleshooting problems), and Interpreting via Visualizations. Two individuals suggested Getting Feedback from Others as a separate activity, but we see it as something that permeates the entire process as we will describe later.

Iterations and Refinements. The iterative nature of the model cannot be emphasized enough. During the collaborative class discussion of the process model, students emphasized the way many activities informed other activities. Sometimes this resulted in completely new analyses, while other times 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., includes 5 minute increments). For example, one student's diary mentions refining her goals several times and collecting 4 rounds of data after visualizing the initial data and viewing other students' work. Other students didn't clearly define their goals until after they had viewed their data. For example, one interviewee said: "*I figured out some way I can collect data and then I just put it into NodeXL to see what would happen and I just kinda got lucky*."

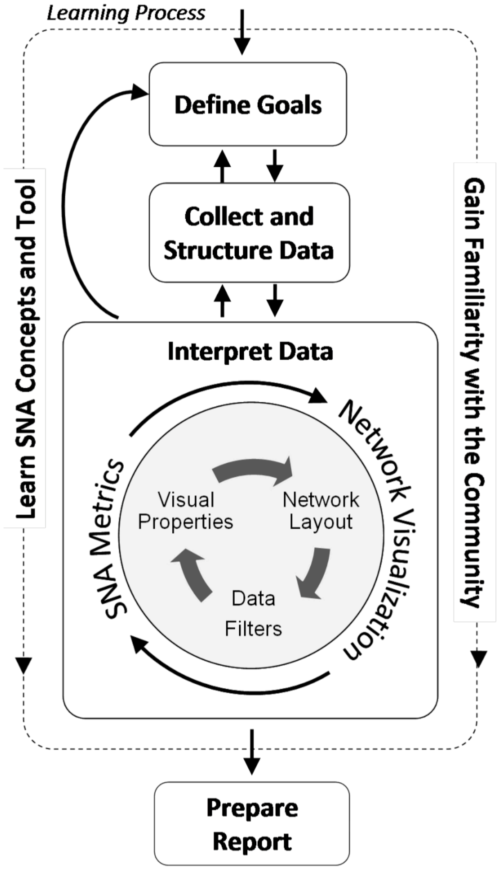

Figure 3. Network Analysis and Visualization Process Model derived from the students' practices.

Although this student had an "aha moment" when viewing the initial dataset, she later collected new data to test some hunches about how one set of wiki pages differed from the original set. Her later reformulation of the problem could not have been developed initially. Another student described how "*a really good research question isn't just stated once and carried through…you collect some data and you go back and you reevaluate your question*."

Impact of Visualization. The role of visualizations was essential throughout the process. In the exploratory cases outlined above, visualization typically provided the key insights that helped the student know if they were on the right track with their data collection and goals. It "*helped put order to a chaotic place like a message board.*"

However, identifying 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 described by another student who said "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 SNA project. We observed several enabling factors and the critical role of collaboration with peers and the instructor. We now discuss these in the context of the major activities shown in the process model.

Define Goals

Most students (8/12) identified defining goals and research questions as the hardest conceptual step in the entire process. When asked what advice students would give others completing the same SNA assignment, 7 of 12 mentioned the need to define specific goals and have some idea of the analyses one will perform before collecting data. The seemingly endless possibilities of analysis given the open-ended nature of the assignment were overwhelming. While the desire to spare other newcomers from the sensemaking struggles around choosing an appropriate goal were heartfelt, doing so may actually deprive newcomers of a necessary step in the learning process. Although most students struggled with this step, a handful of students were able to quickly identify appropriate goals. Indeed, one student mentioned that they were "jealous" of another student who seemed to have such a clear idea of what he wanted to do before even creating his first visualization.

Classroom discussions, instructor consults, and draft ideas posted to the website made it clear that at first several students didn't know what types of questions were amenable to SNA. For example, several students developed questions about correlations (e.g., Do people who post more often also complete more projects?), but it took some time and in some cases expert advice to figure out how to overlay those onto a network structure in a meaningful way (e.g., as is done in Figure 1). Interviews and post-surveys suggest that seeing others' visualizations "opened up the realm of possibilities" for some and was often enough to help someone come up with a "network question." Thus, both expert and peer advice was useful at this stage.

Data Collection & Structuring

The major sub-tasks within this activity included (1) browsing through the site to determine what data could be reasonably collected, (2) deciding what to collect (a task closely related to Define Goals), (3) collecting the data from the community, which occasionally included hand-coding messages into categories of interest, (4) re-structuring the data into an edge list, and (5) adding attribute data into appropriate NodeXL worksheets. Collecting data from communities was primarily a manual process of hand-entering information found on member profiles or other community pages (e.g., discussion forums). This manual process took on average of approximately 2.5 hours as identified in student diaries. This work was described as "tedious" by most, although some data collectors were initially excited with anticipation. In class, several students asked to learn Excel shortcuts and formulas to help collect and format data.

A conceptually challenging aspect of the project was understanding how to structure network data to realize a desired graph. The quiz revealed student confusion over the difference between data stored on the Edges worksheet versus the Vertices worksheet. Four students needed assistance from the instructor to get their data into an appropriate edge list. Additional classroom support for understanding different data structures was required. Students' understanding of the data structures was intimately tied to the visualizations. When discussing how to structure the data, students often did so by describing the desired visualization. The close coupling of data and visualization in the NodeXL application was an asset in supporting the construction of connections between visualizations and their underlying datasets. The visualizations acted as a shorthand for understanding how to structure data, although it required some challenging "network thinking" to understand certain types of relationships (e.g., Figure 2).

On a practical level, some networks were very challenging to represent for non-programmers. For example, the student who created Figure 2 had to manually create the edge list and tie strength values. Creating such data structures manually does not scale well. If a single person has "favorited" 3 fashion designers there are 3 edges to populate (ties between designer A and B, A and C, and B and C). However, if a single person has "favorited" 6 fashion designers there are 15 edges. Another student using this data structure had a friend from computer science create the edge list for her, which required a Perl script. Currently, most network tools (including NodeXL) do not provide features that help users restructure their datasets.

Interpreting Data via SNA Metrics

The major sub-tasks associated with interpreting data via metrics included (1) calculating network metrics within NodeXL, (2) sorting vertices based on the metrics (i.e., to see the individuals with the highest eigenvector centrality), and (3) mapping metrics to visual attributes of the graph.

Students' experience using social network metrics was mixed. Only 1 student visually represented metrics other than *degree* in his final project and 2 others mentioned them in the write-ups despite encouragement from the instructor to include them. This was partially because so many students used bimodal graphs, in which case many of the metrics (e.g., *betweeness centrality*) don't make sense to apply. 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 similar to flashbacks of a Vietnam war vet.

This is one area where peer help was not as useful for students as expert help. This is likely due to these students' unfamiliarity with SNA metrics and the students' lack of confidence in their ability to correctly apply them.

Interpreting Data via Visualizations

The major sub-tasks associated with this step include (1) choosing an initial layout and graph type (undirected or directed), (2) setting visual attributes (e.g., edge width, vertex size, color, shape) to various properties, (3) filtering edges and verices, (4) displaying labels, (5) calculating and viewing clusters, and (6) comparing multiple visualizations.

The role of visualization in interpreting data was critical for students. 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 focusing on visualizations), half the students mentioned "seeing" relationships that might not have been apparent otherwise (e.g., the importance of boundary spanners). For these novices, visualizations were far more important than metrics.

Comments from peers and the instructor on draft visualizations were reported to be helpful. Most students (7/12) thought this was the stage in the process where getting help from peers was most helpful, and 4 thought it was the stage where getting help from an instructor was most helpful. Peer comments on drafts focused primarily on improving visual attributes and layouts, sometimes recommending new variables to map onto visual attributes.

Creating or analyzing visualizations was identified by 11 of the 12 students as the "most rewarding" activity. 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 visual, aesthetic nature of this work mixed with the creative process seems to have resonated 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. Interestingly, some best practices seemed to emerge such as using different shapes to indicate different types of nodes in bipartite graphs (as opposed to using different colors). Although students didn't always know what way of representing a particular attribute was best at first, they often could recognize it when they saw it in others' graphs.

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) writing up the description of the final visualizations and insights learned.

Manually fine-tuning the visualizations took a considerable amount of time. It included sub-tasks of changing colors, changing labels (e.g., shortening them), making visual attributes consistent between multiple graphs, and most time consuming of all adjusting the layout (i.e., positioning of nodes). Most students (7/12) said creating visualizations and in particular adjusting layouts was one of the most technically challenging activities. But it was not considered conceptually challenging. Most students spent over an hour adjusting layouts according to diary entries. The typical approach was to use a built-in layout algorithm (e.g., Fruchterman-Reingold) and then adjust the individual nodes manually. Those who used node labels had a very hard time since they did not want labels to be occluded by the edges.

Overall, students did an effective job interpreting their visualizations. They often relied upon their insider knowledge of their communities to explain why the visualization made sense. Only 1 student (who created Figure 1) included annotations on their final graph, despite the fact that the instructor suggested it as a potentially useful strategy in class. This may have occurred because NodeXL does not allow for annotations of the image directly. Likewise, it does not currently include a legend to link visual attributes such as color to concepts or titles.

One technical challenge that a few people ran into was the lack of support for multiple graphs within the same NodeXL file. A handful of students created visualizations that corresponded with one another in some way, and were frustrated that they had to fine-tune the visual properties of each file manually.

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

Learn and Troubleshoot the SNA Tool

Most students (8/12) mentioned dealing with technology as the task they felt most confused or uncertain about. The usability aspects of the tool are described elsewhere [4]. Here we focus on the role of learning and troubleshooting NodeXL as part of the entire process.

Many students expressed sentiments like the student who said NodeXL "is very strong, very nuanced, but not very approachable." She later said she felt like she had completed one cooking class and was given a set of very sharp knives to use. This is likely a problem primarily because students didn't have a clear mental model of network analysis itself, which led to confusions with the technology such as understanding what data should be on the vertices tab versus the edges tab and interpreting concepts like “tie strength”. Having many different ways to accomplish the same task also confused students at times.

Many students (7/12) said getting help from the instructor was most helpful when learning Excel skills and NodeXL features. Novice Excel users had trouble using formulas, entering data, and didn’t use time-saving shortcuts like autofill. Only 2 students identified peers as being helpful in learning the technology, but observations of the lab session made clear that students helped each other more often than they remembered. The resource most used throughout was the tutorial, which was deemed helpful by students, although some complained that the datasets used in it were too simple making translating the ideas to their communities challenging.

Another challenge students faced was dealing with error messages and sporadic software. Because only an early version of NodeXL was available when the class took place, 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”).

DISCUSSION

Process Model Comparison

The process model derived from the student experiences (Figure 3) fits well within the larger set of models that describe sensemaking activities [18, 22], particularly those that consider visual analysis [15]. Our model was derived from the experience of novices rather than expert intelligence analysts. As such, it is not a model of SNA best practices that would emerge from extensive experience. Therefore, it should not be used prescriptively without caution. However, the apparent success of the students suggests that adopting such a model may be beneficial for those new to network analysis. Furthermore, observing novices helped us develop insights into the challenges they face and potential educational and design aids for them.

Like Kulthau’s model of information seeking [14], we have considered an affective dimension. The strong feelings identified from user diaries often reflect the struggle of novices with an unfamiliar sensemaking process. Feelings of uncertainty, for example, accompanied unclear and underspecified user goals while increased interest and excitement were often associated with the initial viewing of data visualizations and the sense of satisfaction at the completion of a task or a project. With this model, instructors may help students set accurate expectations about the process. The affective dimension also suggests that NodeXL should be modified to help reduce the feeling of being “overwhelmed.” Improving design layouts and adding an ‘undo’ capability would help.

On a substantive level, our model is consistent with other sensemaking models, though its strong focus on visualization led to some unique observations. Indeed, similarly to [18] we see a progression in the user’s activities, starting with the data collection, leading to the formulation of more structured data, and then models, in our case visualizations, and a final report. They also observed the iterative nature of this process, where analysts moved back and forth between different phases.

More than existing models, our model emphasized that new SNA analysts relied heavily on *visualization as the means of sensemaking.* The model presented by [15] can be used to shed more light on this finding, although it did not solely focus on visualizations. Klein et al. describe the process by which users create a *frame* that describes data, elaborate the frame, and compare alternate frames leading to modifications or adoption of new frames [15]. In our study, visualizations were created to explain community data, thus serving as a frame to manage attention, and define, connect, and filter the data [15]. We observed that students had trouble creating new frames (i.e., data structures and graph types), but they were effective at adapting frames they observed from others to their communities. The elaboration process was less pronounced in our study, overshadowed by the desire to refine existing frames to make them more understandable to others. Visualizations were used by students to generate discussion around communities and succinctly transfer insight to others. Visualizations also served as the foundation of the new language of SNA that students learned and shared.

Implications for Designers

The study led to numerous design suggestions, some of which are specific to the NodeXL tool and reported on elsewhere [4]. Here we focus primarily on designs that enable *collaborative learning and use of SNA in practice* and creating effective visualizations for data *exploration and communication of findings* by community analysts.

Support for collaboration. Currently, NodeXL and other SNA tools are stand-alone programs with few built-in collaborative features. Our study suggests that both peer and expert advice can be enormously helpful at several steps in the process. This has implications on the support for exchange of NodeXL files and graph images. The current implementation is far from ideal since each NodeXL installation renders images differently depending on the default graph settings. Exchanging images would be more effective if they could include an automatically generated legend that mapped data variable descriptions such as column names to visual attributes, e.g., color, size, and shape. Adding summary statistics, e.g., those already calculated in NodeXL, would help provide additional context and encapsulate related information.

Furthermore, the collaborative student practices point to a need for a community of SNA visual analysts as a forum for practitioners to exchange experiences, provide feedback, and receive advice, in some aspects similar to ManyEyes [23]. Enabling the users of SNA tools to create composite objects of visualization and context metadata and upload them in a simple step, would connect them with the community. The service would enable download and remix of data with attributions provided as in other collaborative communities like Scratch [19]. The repository would be enhanced by classifying visualizations based on similar data structures (e.g., bimodal, directed) and topics (e.g., wiki, forum, proteins), as well as rating quality (e.g., using NetViz Nirvana quality metrics [10]). Facilitating communication via comments and annotations would likely foster meaningful exchanges like those we observed.

Balancing the user effort. Tools like NodeXL have removed the complexity of programming algorithms and metrics for network analysis. These operations can now be achieved by a simple click or selection of NodeXL functions and lead to a rewarding experience of visual exploration and iterative refinement. However, two stages in the process deserve significant attention by designers: (1) the transformation of original datasets into alternate data structures (e.g., implicit graphs like Figure 2), and (2) the report preparation phase that still requires significant manual positioning of nodes to reach NetViz Nirvana.

The iterative nature of SNA exploration suggests the need for tighter integration of source data (e.g., forum posts) and visualization. For example, the analyst could use the network representation to navigate, filter, and search through the data points and vice versa, the same operations on the raw data can result in the dynamic changes of the network representation. Currently, the NodeXL data representation is extendible, allowing new labels and URLs to be added to a right-click menu for each object in the graph pane. However, more effective zoom from graph abstractions to the data would be desirable.

Finally, the study shows that visual aspects of graphs played an important role in processing the data. Therefore it is important to provide a spectrum of data layouts that are easy to apply and understand by the user. Observations of the students' practices indicate that default parameters for the layout are critical in order to avoid tedious manual intervention in order to achieve quality graph. Including visual analytic metrics into the tool as is done in Social Action [17] would encourage the creation of more understandable graphs by those without training.

Thus, providing effective data converters for preparing edge lists from source data, better layout algorithms with clear visual analytic measures, and easy access to data would eliminate a significant overhead for the user. It will establish a balanced process model for SNA with a better ratio of effort involved in the data preparation, access, and visual analysis.

Implications for Educators

Structuring SNA Assignment. The SNA assignment students completed worked well in forcing students to consider the entire SNA and visualization process. Having students come up with their own goals was conceptually challenging, yet a useful learning exercise. Figuring out what data to collect and how to structure it was also a vital part of learning to think in terms of networks. However, the tedious process of collecting data manually for several hours did not have any direct payoffs. A better approach is to allow students to use one of a few pre-prepared datasets (e.g., based on Wikipedia or Twitter data) with numerous variables that students would need to figure out how to apply. Motivated students could also be given the option of collecting their own community data.

Facilitating Collaboration. Facilitating peer collaboration more actively would have also been beneficial. Several students complained that there was not enough time to informally play around with the tool, particularly in group settings, since the first assignment was to perform an individual quiz. Students who attended the optional lab session recommended setting aside more class time for similar group study time. Commenting on each others' visualizations, even remotely through the course website was successful both for the value of the comments and even more for providing students with fresh ideas from seeing others' work.

Implications for Online Community Analysts

SNA's Promise. We believe this study provides sufficient evidence for the powerful and unique benefits that network visualization and metrics afford community analysts. Just as the map visualization in Google Analytics provides a quick visual overview of website traffic by country, a network visualization of community interaction provides an overview of social interactions within a community and the basis for detailed analysis. SNA is particularly useful when trying to identify unique individuals, subgroups, and relationships between them. For example, SNA metrics such as *betweeness centrality* and visualizations such as Figure 1 can identify boundary spanners, who are important not because of the number of posts but because of their unique position within the network. As these tools become more widely used, they will enable more systematic tracking and comparisons of communities helping analysts more effectively create sustainable, engaging interactions. Furthermore, community visualizations can be excellent conversation starters, tell important stories about the nature of your community and changes that have occurred, and help community members be a bit more self-reflective.

Accessibility of SNA. 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 backgrounds were able to effectively apply SNA to analyze online interactions.

CONCLUSION

Since social media tools may have a profound transformative effect on society, software tools and systematic processes to analyze complex social interactions will play a key role in raising their efficacy in communities and organizations. Our focus has been on developing software tools and intellectual processes that facilitate social network analysis by a wide range of users, not just academics and statistical professionals. To validate and refine our tool (NodeXL) and process model (Network Analysis and Visualization Process Model), we studied 15 graduate students during 5 weeks of intensive learning of the tool and process model. Their struggles and innovative successes demonstrate that non-technical students can acquire skills necessary to conduct worthwhile analyses of social media usage. The detailed empirical study led to numerous suggestions for fixes and improvements to NodeXL, many of which have already been made. The rich implications for educators are being embedded in our current revision of teaching materials. Our implications for community & organizational analysts will be more difficult to test, but upcoming tutorials and trials with professionals will provide valuable case studies.

ACKNOWLEDGMENTS

REFERENCES

1. Adar, E. and Adamic, L. Tracking information epidemics in blogspace. *Proc. Int. Conference on Web Intelligence 2005,* ACM Press (2005), 207-214.
2. Barabasi, A. *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, NY, 2003.
3. Beyer, H., and Holtzblatt, K. *Contextual Design: A Customer-Centered Approach to Systems Designs*. Morgan Kaufmann, San Francisco, CA, USA, 1997.
4. 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. Submitted, 2009.
5. Brown, B. A., Sellen, A. J., and O'Hara, K. P. A diary study of information capture in working life. In *Proc. CHI '00*. ACM Press (2000), 438-445.
6. Burt, R. S. *Brokerage and Closure: An Introduction to Social Capital*. Oxford University Press, New York, USA, 2007.
7. Carter, S., and Mankoff, J. When participants do the capturing: the role of media in diary studies. In *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 Publications Inc, Thousand Oaks, CA, 2007.
9. Cross, R., Parker, A. and Cross, R. *The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations*. 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. C. *The Development of Social Network Analysis: A Study in the Sociology of Science.* Empirical Press, Vancouver, BC, 2004.
12. Google Analytics: http://www.google.com/analytics.
13. Hansen, D., Shneiderman, B., & Smith, M. *Network Analysis with NodeXL: Learning by Doing*. Unpublished manuscript.
14. Kuhlthau, C. C. *Seeking meaning: A process approach to library and information services*. Libraries Unlimited, Westport, CT, 2004.
15. Klein, G., Phillips, J. K., Rall, E. L. and Peluso, D. A. A Data-Frame Theory of Sensemaking. In *Proc. of the Sixth International Conference on Naturalistic Decision Making*. CRC Press, (2007), pp. 113-155.
16. Lithium Technologies. Community Health Index for Online Communities. White Paper, 2009.
17. Perer, A., & Shneiderman, B. Systematic yet flexible discovery: guiding domain experts through exploratory data analysis. In *Proc. IUI '08*, ACM Press (2008), 109-118.
18. Pirolli P., & Card, S. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Through Cognitive Task Analysis. Paper presented at *International Conference on Intelligence Analysis*, 2005.
19. Scratch: http://scratch.mit.edu/.
20. Smith, M., Shneiderman, B., Milic-Frayling, N., Rodrigues, E. M., Barash, V., Dunne, C., et al. Analyzing Social (Media) Network Data with NodeXL. Forthcoming In *Proc. C&T 2009*.
21. Telligent's Harvest^{TM} reporting server: http://telligent.com/products/harvest-reporting-server/.
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. Wattenberg, M., Kriss, J., and McKeon, M. ManyEyes: a Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics 13,* 6 (2007), 1121-1128.
24. Welser, H., Gleave, H., Fisher, D., and Smith, M. Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure, 8*,2, (2007)