Digital Library Framework for Behavioral Science

Prashant Chandrasekar
Virginia Polytechnic Institute and
State University
Blacksburg, VA 24061 USA
peecee@vt.edu

ABSTRACT

Much of the success in a user study, that analyzes the behavior of people in a social network, is based on how the study (or in this case, a "social-network-based" study) is conducted and how outcomes are measured. It is challenging for behavioral scientists (e.g., psychologists and sociologists) to set up a controlled user study on a social-network, and analyze the data from that study. We describe our plan to build a digital library framework that would significantly aid researchers to plan and execute such comprehensive user studies. We identify key components that are required to conduct a successful social-network-based behavioral study, types of data and metadata that are critical to such studies, and methods to fuse the heterogeneous mix of information involved. We discuss services of the digital library, that measure outcomes (such as engagement) of user studies, and aid analysis, e.g., of textual content. Through the digital library, researchers will have an end-to-end template on how to design and implement a social-network-based user study, along with services that would enable them to analyze and measure outcomes for hypothesis testing.

Keywords

Digital Libraries; Data-Mining; Social Networks; User Studies;
5S Framework

1. BACKGROUND

This research, which is a central part of a doctoral dissertation, is motivated by the Social Interactome [1], an NIH-funded study involving the Addiction Recovery Research Center at VTCRI and the Department of Computer Science at Virginia Tech. The project aims to identify the various scenarios and conditions in which people, in recovery from substance addiction, when placed in a constrained network topology and provided with self-management tools and other aids, can advance along their road to recovery. The study explores ways to select a *helpful* set of "recovery buddies" who will provide social network support for recovery. One hypothesis is regarding network topology, i.e., that selecting sets of recovery buddies with maximal overlap, as opposed to choosing randomly, will be more helpful. Another hypothesis is that homophily-based sets of recovery buddies will be more helpful. *Helpfulness* is measured to reflect aiding greater engagement and a more sustained recovery. We are running several sets of experiments, with replicates ensuring generality. Each replicate is a clinical trial, where 256 participants, each of whom is recovering from addiction, are recruited and placed in one of two social networks containing 128 participants each. One network is for the treatment; the other is a control group. Results will be compared between the two networks. Experiment 1 has been completed. Two replicates each ran for 16 weeks, with some participants continuing longer.

2. MOTIVATION

The Social Interactome is the first randomized clinical trial of online social networks to support recovery from addiction, additionally, with the utilization of evidence-based interventions such as the Therapeutic Education System (TES) and other self-management tools. From a computer science perspective, we have a unique opportunity to study data indicating the effects of the various interventions, along with data from website access logs, textual communications, and self-reports.

Furthermore, through the various types of data that we collect from all of the experiments (and replicates), we can identify successful practices for similar user studies; design, build, and train models to define outcome variables such as engagement, social influence, psychological state, etc.; and through observation and learning, build dependency graphs that can help identify causal effects that result in progress towards recovery.

We spent over a year carefully designing every aspect of the social network and information system to support the analysis of this study. Through the course of the first experiment, we have built a set of tools, database models, and scripts to test hypotheses of the study.

Through this process, we have had a unique opportunity to observe the overall needs of the key investigators of the study, learn the types of analyses that they are looking to do to validate their hypotheses, and discover a variety of psychology-related documents that discuss measures used in behavioral studies.

It takes a great deal of effort and expertise to plan and carry out such studies. Much additional effort is needed for each replicate and each subsequent experiment in the same overall study. Since the efforts after the first experiment require much of the same type of analysis, automation could save the time of researchers and lower the cost of such multi-experiment investigations. A digital library framework could encompass key facets of the process such as identifying informative data points, computing measures, building models, and testing/validating hypotheses. We aim to provide such backbone support, that could aid, if not drive, future psychology/sociology-related studies.

3. RELATED WORK

One of the main focuses for the NIH funded grant, and an issue in the planned Ph.D. research, is understanding the effects of network topology on information flow and website engagement. Dr. Centola, a consultant for the project, has conducted many experiments on synthetic networks and has hypothesized positive benefits from highly clustered networks with homophily (considering demographics), among other features. In our research, we evaluate his findings, appropriately including these factors into our models and software [5][6][7].

Many studies confirm the importance of social groups in influencing a variety of behaviors such as alcohol consumption, happiness, obesity, etc. These studies have considered and aimed to model social influence and their role regarding human behavior [8][9][10][11].

Work conducted by a team at the University at Oregon deals with spreading wellness and predicting the behavior of people in a social network who are each suffering from obesity. They model social influence and predict activity and wellness using self-reporting of personal fitness, and infer social influence by finding correlations in similar activity between friends. Our work differs in that the setting for our network is slightly more constrained and we are attempting to understand psychological impacts on participants through our study using not only the self-reporting of substance usage, but also social activity in the network (e.g., website use and communication between participants) [12][13].

The 5S model provides us with a way to formally describe a digital library using five abstractions: streams, structures, spaces, scenarios, and societies. They allow us to identify and describe/represent concepts such as services, digital objects, metadata, and collections that serve as foundations for the components and information flow in a digital library [14][15]. Streams represent the flows of information (for example streams of characters representing text) that are collected and stored in different file formats, each of which represents a different Structure (such as XML or JSON). Each of the documents and collections go through various transformations, e.g., being represented in a variety of Spaces (such as "feature spaces" or "vector spaces"). Writing Scenarios of system use is critical in understanding the workflow, procedures, etc. that help a user achieve a task. Scenarios represent the use of the digital library and tell us what happens to the stream of information, in the spaces, through the structures. Taken together the scenarios describe services, activities, and tasks—and those ultimately specify the functionalities of a digital library. The Society served includes the main group of individuals for whom this library is built, i.e., researchers from various disciplines who are involved in studying behavior in social networks.

One of the goals and functions that the digital library should serve is to allow for these behavioral studies to be re-run and possibly extended for further research. Therefore, our system would need to support research reproducibility through simulation. The CINET system, built by the NDSSL group at VT, has been designed to encourage and support network science research. One of the many services provided by the system is the modeling of networks and simulation of analyses. This system employs a "simulation-supporting" digital library to aid network research [19]. Along the same lines, SimDL, developed by Jonathan Leidig, is a digital library framework that maintains datasets, input configurations, results and other related documents, that helps streamline the process of running simulations of various studies in computational epidemiology [20].

4. PROBLEM

The set of problems that we have identified and base our research on, are as follows:

1. *There is no digital library framework and/or system that is tailored to support research on behavioral studies in a social network.*

2. *It is extremely cumbersome, among and across related research studies (e.g., different clinical trials), to replicate/expand on the study of existing statistical models for social-network-related outcomes variables.*

3. *It is difficult to re-purpose and/or extend a social-network-based behavioral study, to be applied to different contexts with distinct data sets.*

4. *There is no digital-library-based formal approach or framework that represents the key features in behavioral studies, that could be instrumented across different experiments/studies.*

More specifically regarding the design of the digital library, some of the problems to consider are:

5. *How should we represent all the data, models, and toolkits/frameworks as services that could be easily extended by future psychology/sociology researchers?*

6. *How can the 5S framework for digital libraries help with description, design, implementation, and evaluation aspects of similar digital libraries aimed to aid psychology/sociology researchers working with participants involved in social networks?*

5. DATA

We have been collecting and organizing data from each of the participants for all of the experiments. Replicate 1 of the study commenced October 27, 2015 while replicate 2 began April 11, 2016. Each replicate of experiment 1 ran for 16 weeks (plus some participants continued), with 128 participants placed in two different networks. During the first 12 weeks of the study, each participant could interact with 6 other participants, plus the administrator. From week 13, the social network was opened up so that participants have the opportunity to discover and add more people to their recovery buddy network.

The data from the study fits broadly into three categories:

1) Web logs: We have instrumented web analytics into the study through the open-source framework, Piwik [2]. Through this framework, we have a detailed history of the "clicks" of the participants, during website visits. This includes when they logged in, every page visited, and the length of each of their online sessions.

2) Psychology-related data: The social network interface has components that allow participants to complete surveys that request information about their daily substance use and their social group; additional questions have been created by the psychology wing of the team. These aim to obtain information deemed important to better understand the circumstances of each participant. Additionally, participants have opportunity to attend "Video Meetings." These are online group sessions with discussion led by a certified moderator. The website also includes testimonials from peoples' experiences with recovery. These testimonials are taken directly from the International Quit and Recovery Registry (IQRR) website [4]. Finally, the participants have the opportunity to view a wide range of "self-help" and/or "self-management"-based modules in our Therapeutic Educational System.

3) Social-network-related information: During the study, the participants have opportunity to communicate with each other through posts, shares, likes, etc. The interaction can be in textual form and/or through sharing of photos, web links, videos, etc.

The surveys, mentioned above, generally fit into three categories:

- Weekly Assessment: Once a week, each participant has the opportunity to report on their substance usage for the week.
- Time-point Assessment: Three times during the course of the 16 weeks of the experiment, the participants can answer questions that surround and pertain to their background, as well as current circumstances of their recovery from addiction.
- General Assessment: These sets of assessments are a part of the International Quit and Recovery Registry [4], as well as of the "self-management" tools that they can use for personal benefit.

The participants are monetarily compensated for completing the weekly and time-point assessments.

The main goal of the experiment is to study how website engagement and peer engagement, both of which are measured through web logs and social-network-related information, impact the progress of a participant's recovery (measured through the psychology-related information). More precisely, we are investigating if (and how) social engagement (or activity) effects each participant's substance use over the period of the study.

The research team identifies social engagement and participant relapse as the key outcome measures for this study. Our system will support development of statistical models to measure these outcomes. Key components of the library support analysis of these measures. Measures derived from web-logs and social-network-related information are common among studies involving social network data. Therefore, the analysis of these measures is applicable across studies. However, analysis of participant substance use is an outcome measure that pertains to this study alone. Therefore, one of the challenges (and goals) of the study is to build a system that has the capability to evaluate any and all measures, based on the user's data and the outcome measure desired. This includes, but is not limited to, having a generalized input data schema so processed information is fed into our models to measure the outcome variables.

6. RESEARCH METHODOLOGY

The Social Interactome represents a baseline or "use case" for behavioral studies that our digital library would support. Therefore, the initial design of the components of our digital library would be a generalization of what supports the current experiment. From a top-down approach, we first identify the various scenarios in which researchers interact with the digital library, based on the data and analysis requirements specified by the co-PIs and other researchers in the project. These scenarios would naturally translate to services that the digital library should offer. The analysis that pertains to the hypothesis testing specified in the grant would be generalized using a repository of models to be provided through services. Each of the models will be testing a hypothesis based on measures that we define and infer from the raw data (which in turn are tied to scenarios we are discovering that are carried out by participants in the experiment). We will be defining a set of such data transformations to clean the raw data and extract measures from it. This work can be guided by the 5S framework.

The different components will be expanded/refined as we identify more scenarios, which leads to implementation of more services, that would require building of additional models, based on newly identified/defined measures from the data (and underlying participant scenarios and actions).

As an example, we describe below aspects of the approach to design the digital library. Thus, we identify key components of the library that would serve a generalized version of the user requirements.

1. Scenarios (observed from the researchers):

1.1 The research team wants a set of reports, produced on a weekly basis, that compare the participant activity between both network topologies.

1.2 The moderator wants to be alerted if/when a participant is in distress, so that they can monitor the situation and act accordingly.

1.3 The research team wants to know the percentage of people who have completed the "paid" assessments across both network topologies.

1.4 The research team wants to observe if one of their hypotheses, i.e., ***"There is a positive correlation between participant engagement in the social network / website, and recovery from reported substance use,"*** is being validated.

2. Services (to be provided to support the scenarios):

2.1 Participant Activity Report: An API to request aggregate numbers indicating user activity (for each feature of the website as well as any self-reported substance relapse) for both networks, given a time range.

2.2 Distress Prediction: An API to analyze a set of texts in files, provided as a parameter, to determine if any text is indicative of distress.

2.3 Same API call as 2.1, but in this case, selecting a particular feature (the activity on paid assessments), to generate a report on use.

2.4: Engagement Score Report: An API to request a score given to each participant based on statistical models that measure engagement. The substance relapse information can be retrieved using service 2.1.

3 Models: Based on the data requirements, we have focused on building the following models that we believe apply to social-network-related research.

3.1 Website Engagement: A model to determine if, and how, participants are engaged in the study.

3.2 Social Influence: A model to characterize (and provide a score for) social influence between pairs of participants.

3.3 NLP-related models: A set of models that involve text processing, as well as unsupervised and supervised machine learning methods, to identify: a) whether a participant is expressing distress and might be in danger, b) whether they are sharing a milestone or success in their road to recovery, and c) topics of discussion between participants.

4 Measures: For this particular project, we had initially identified a list of measures that could be important variables for a variety of models. Some general measures that apply to such studies are:

4.1 Usage Metrics: # of Logins/Day, # of pages visited, total time spent on website, frequency of logins to the website, etc.

4.2 User Inferred Metrics: Sentiment of each post, pairwise strength of communication between participants, etc.

4.3 Communications: # of posts on a wall, # of messages sent, # of comments/likes/shares.

7. PRELIMINARY WORK (Social Interactome)

This section details some of the efforts that have been applied to the experimentation, and in setting up and executing preliminary analysis of the data. We conducted exploratory analysis to understand how participants behave in such a forum.

7.1 Development Work

The Department of Computer Science personnel, along with developers from the Addiction Recovery Research Center (ARRC), designed and built the infrastructure and interface to conduct the first experiment. Friendica [3], an open source community social network development project, is the foundation for our system. We modified Friendica so we could establish network constraints as per study needs, and integrated the various self-management surveys and tools into the interface. To track every "web-click" of each participant, we instrumented Piwik [2] into the code. Our operational environment consists of a production server (where the website is hosted) and an "analytics" server, where all of the data is captured and stored in its raw form for analysis.

7.2 Text Analysis

Prior to the start of the experiment, there were two particular areas on which we wanted to focus:

1) We should be able to detect if a participant is expressing distress. Since the study is with live participants, we wanted to ensure that we are alerted if/when a participant expresses suicidal intent. It is not feasible for a moderator to read each and every text created by any of 256 participants in real-time. Therefore, we built a module to parse texts as they are created and alert the team if certain "keywords" are present. The members of the ARRC team helped build the dictionary of terms that would signal "distress". The dictionary was updated as we came across additional terms from texts created in the experiment, or from other sources.

2) Explore whether and how participants express any small success they achieve in their road to recovery: as mentioned in the research methodology, much of the initial analysis was of an exploratory nature, involving observing how people express themselves when online. Over the course of the experiment, we gained a better understanding of the vocabulary and language used in such forums. Our initial hypothesis was that the textual content would be along the lines of testimonials/success stories found in websites that are related to recovery from addiction. One representative excerpt from a success story is:

"January 2013, was my turning point. I stop using drugs, I stopped drinking & that was all I was doing"

It was equally important that we also learn how people wrote about the low points in their life. Therefore, along with a group of undergraduate researchers, we built a text classifier that would predict if a text is expressing "success." We decided to use the set of success stories as part of a training corpus. Success stories and/or testimonials on sites such as IQRR [4] often include the entire journey of the author. We extracted sentences that pertain to either of the labels (success or not) and built the training set. A bag-of-words model was used and a multinomial implementation of the Naïve Bayes classifier was applied to the text.

7.3 Constructing Participant Profiles

We are currently working towards building a feature vector as a representation of each of the participant's "profile". The features include demographic information along with psychology and substance-abuse-related information. As a requisite to qualify for the experiment, each participant completes an "enrollment" survey that requests such information. Some questions in the survey help derive the personality trait of the participant through the standard Big Five Personality Traits model [16][17]. We also collect lifetime substance abuse and relapse information on all the substances that they have used and are recovering from. In addition, as part of the time-point assessments, participants provide information to help derive their substance abuse dependence as defined by the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, 5th Edition) criteria [18]. Using the criteria, a participant could either be mildly, moderately, or severely dependent on a substance. Finally, through the weekly assessments, we collect the relapse information used in a 16-week summary of progress towards recovery. Each of the above types of information serve as attributes or characteristics of a participant. We plan to combine this information in many ways to create logical groupings of participants that are similar, and analyze their website usage and textual information as a group to identify possible correlations in the data.

7.4 Understanding Success/Failure of Study

During the study, participants can engage in the various features of the website. The design of these features was based on providing a supportive environment for participants to share experiences with one another, get inspired by reading testimonials from others who have made a promising recovery, learn through various educational modules, and participate in virtual group meetings. The surveys act as self-management tools, and, along with other system features, aim to aid each of the participants in their journey towards recovery.

The participants in the study are monetarily incentivized to complete the weekly assessments and time-point assessments. The rest of the interactions on the website are not so incentivized. Our hope is that the participants are motivated to engage in the not incentivized features of the website either through moderator interventions, that include icebreakers to help introduce participants to one another, or posts to encourage interaction on the website, or by being encouraged/influenced by their recovery buddies on the website.

First, we analyze, at an aggregate level, the behavior of participants towards the incentivized features of the website versus the (optional) features for which they are not being paid. Among the features that are "optional", we would like to identify the ones that were most popular, and study if this engagement propagated through the social groups. Next, the TES modules and the stories can be categorized according to substance; we assume they would be more relevant to participants recovering from one substance over another. For example, there are stories and TES modules that are specific to alcohol addiction. We would like to investigate any correlation between the topic/content of these stories and modules and the primary substance addiction as reported by the participants. We would further look for patterns based on other features/characteristics of the participants. We also plan to analyze the effectiveness of moderator posts based on response from participants, via comments, likes, and shares, and to see if a post from the moderator has started a conversation between participants and their recovery buddies. For this, we would analyze the text generated in the period shortly after the moderator post, to check for convergence in language or topic.

7.5 Reports

In an effort to distinguish signal from noise, we generated reports to aid understanding of how the interface was being used, as well as to investigate interactions between participants. Figure 1 shows the number of texts created in each of replica 1 of experiment 1's two networks.

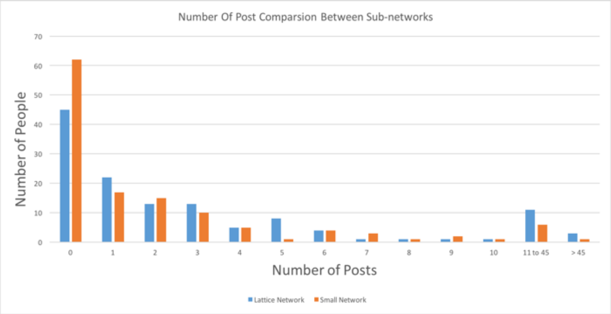

Figure 1: Comparison between the two sub-networks in the number of textual posts broken down by the number of people

The social network has a feature that allows users to privately message one another. We understand that this feature can play an important role in helping build relationships/friendships among participants. Figures 2 and 3 illustrate the private messages sent across the networks. The vertices represent the users, while edges represent communication between users. We observe that more participants in the lattice network (with overlapping sets of recovery buddies) seem to have used the private message feature. In the small world network (with randomly chosen recovery buddies), however, most of the private message communications have been to the moderator of the experiment (shown as the vertex at the center of the graph).

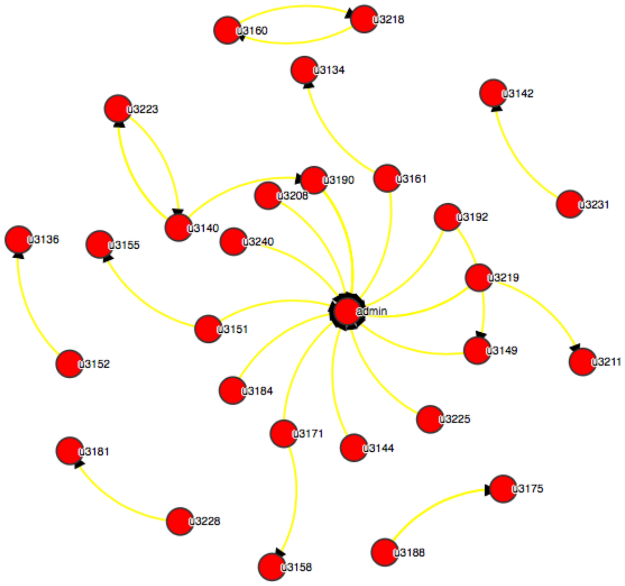

Figure 2: A graph that represents communication between users in the small world network. Many of the edges are connected to the node at the center, i.e., the moderator of the experiment.

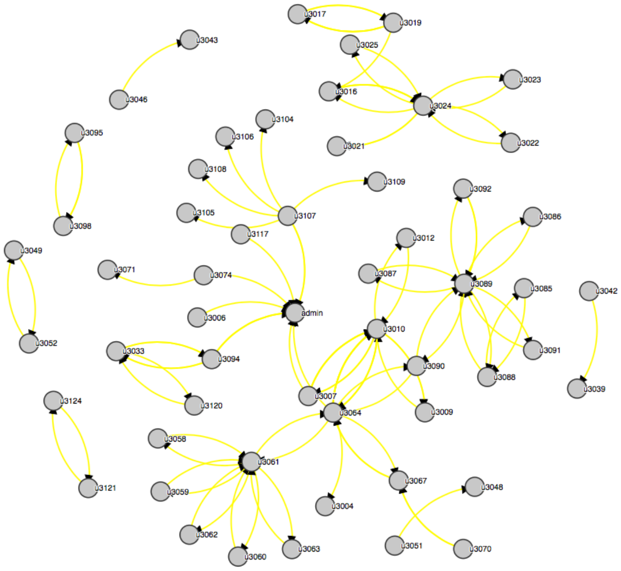

Figure 3: A graph that represents communication between users in the lattice network. This illustrates how having overlapping sets of recovery buddies can lead to more and larger connected components.

8. CONCLUSION

There are many considerations when designing a full-fledged behavioral study, especially when involving a social network. The success/failure of a study is based on how each of them is addressed. Through our digital library framework and implementation, we provide researchers with guidelines to support user engagement, effective user interventions, data capturing mechanisms, and models to analyze the data from the studies. We hope that a generalized DL, which would be the outcome in the planned Ph.D. research, will not only aid in automating the early work done for the Social Interactome during the later experiments, but also provide a framework for other psychologists/sociologists to create and test various hypotheses that relate to the study of human behavior in social networks.

9. ACKNOWLEDGEMENTS

This research is based on the work towards the NIH-funded project, The Social Interactome of Recovery: Social Media as Therapy Development (NIH Grant 1R01DA039456-01).

10. REFERENCES

[1] The Social Interactome, https://quitandrecovery.org/the-social-interactome/, accessed on 2015/02/08

[2] Piwik Open Analytics Platform, http://piwik.org/, accessed on 2015/02/08

[3] Friendica, Open Social Network, http://friendica.com/, accessed on 2014/10/12

[4] International Quit & Recovery Registry, https://quitandrecovery.org/, accessed on 2014/10/12

[5] Centola D. The spread of behavior in an online social network experiment. Science. Sep 3, 2010; 329(5996):1194-1197

[6] Centola D. Are health behavior change interventions that use online social networks effective? A systematic review. 2013; 127:2135-2144

[7] Centola D. An experimental study of homophily in the adoption of health behavior. Science. Dec 2, 2011; 334(6060):1269-1272.

[8] Rosenquist JN, Fowler JH, Christakis NA. Social network determinants of depression. Molecular psychiatry. Mar 2011; 16(3):273-281.

[9] Rosenquist JN, Murabito J, Fowler JH, Christakis NA. The spread of alcohol consumption behavior in a large social network. Annals of Internal Medicine. Apr 6 2010; 152(7):426-433, W141

[10] Fowler JH, Christakis NA. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. British Journal of Medicine (Clinical Research Ed.). 2008, vol. 337/no., pp. a2338-a2338.

[11] Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. The New England Journal of Medicine. Jul 26, 2007; 357(4):370-379

[12] Yelong Shen, Ruoming Jin, Dejing Dou, Nafisa Afrin Chowdhury, Junfeng Sun, Brigitte Piniewski, and David Kil. "Socialized Gaussian Process Model for Human Behavior Prediction in a Health Social Network." In Proceedings of the 12th IEEE International Conference on Data Mining (ICDM 2012). pp. 1110-1115, 2012

[13] Shen, Y., Phan, N., Xiao, X. et al. Knowl Inf Syst (2016) 49: 455. doi:10.1007/s10115-015-0910-z

[14] M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. ACM Transactions on Information Systems, 22:270–312, April 2004

[15] Edward A. Fox, Marcos Andre Goncalves, and Rao Shen. Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach. Morgan & Claypool Publishers, San Francisco, July 2012, http://dx.doi.org/10.2200/S00434ED1V01Y201207ICR022

[16] Goldberg, L. R. (1993). The structure of phenotypic personality traits. American Psychologist 48: 26–34.

[17] Costa, P.T., Jr. & McCrae, R.R. (1992). Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) manual. Odessa, FL: Psychological Assessment Resources.

[18] American Psychiatric Association. DSM-5 Task Force, American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th ed. Arlington, VA: American Psychiatric Association; 2013

[19] Abdelhamid, S. E., R. Alo, S. M. Arifuzzaman, et al. "CINET: A Cyberinfrastructure for Network Science", 2012 IEEE 8th International Conference on E-Science, (2012), pp. 1-8

[20] Jonathan P. Leidig, Epidemiology Experimentation and Simulation Management through Scientific Digital Libraries; Virginia Tech Department of Computer Science Doctoral Dissertation, 2012