ABSTRACT

Persons who engage in non-suicidal self-injury (NSSI), often conceal their practices which limits the examination and understanding of those who engage in NSSI. The goal of this research is to utilize public online social networks (namely, in LiveJournal, a major blogging network) to observe the NSSI populations communication in a naturally occurring setting. Specifically, LiveJournal users can publicly declare their interests. We collected the self-declared interests of 22,000 users who are members of or participate in 43 NSSI-related communities. We extracted a bimodal socio-semantic network of users and interests based on their similarity. The semantic subnetwork of interests contains NSSI terms (such as “self-injury” and “razors”), references to music performers (such as “Nine Inch Nails”), and general daily life and creativity related terms (such as “poetry” and “boys”). Assuming users are genuine in their declarations, the words reveal distinct patterns of interest and may signal keys to NSSI.

Author Keywords
Interest, NSSI, Self-Injury, Similarity, LiveJournal

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—Natural Language

General Terms
Experimentation, Human Factors, Languages

INTRODUCTION

Non-suicidal self-injury (NSSI) is the direct, deliberate destruction of one’s own body tissue in the absence of suicidal intent [9]. It is practiced primarily by adolescents and young adults [3] and is often concealed from others. Common NSSI activities include skin cutting, banging or hitting oneself, and burns [4].

Recent prevalence estimates suggest that 14% to 21% of adolescents and 17% to 25% of young adults have engaged in NSSI at some point in their lives [4, 10]. Furthermore, NSSI is repeatedly found to be associated with significant emotional and behavioral dysfunction (e.g., eating disorders, suicide [8]). These findings highlight the need to enhance understanding and prevention of NSSI and its psychiatric sequelae.

The goal of this research is to find mechanisms that could identify NSSI persons by automatically analyzing secondary data publicly available from massive online social networks (MOSN), without explicitly interacting with the subjects.

Many popular MOSNs (e.g., Facebook and LiveJournal) allow users to declare their interests, either explicitly or in the form of “likes.” While these interests are often selected randomly and polluted with “status words,” we found a very significant correlation between interest lists and membership in NSSI online communities in at least one major MOSN—LiveJournal [7, 12], a blogging social network. We used the declared interests as nodes and similarities between their users as edges to build a semantic network. The layout of the network consists of four clearly separated word clusters, one of which corresponds to the pathological terms (e.g. “self-injury” and “razor”) and the other three refer to daily life, popular music, and creativity. We expect that the bridge terms that connect the pathology cluster with the remaining three clusters can be used as beacons signaling the potential presence of an NSSI behavior.

The rest of the paper is organized as follows: in Section 1, we describe the data acquisition process; Sections 2 and 3 explain the semantic network generation and the resulting network organization; network comparative assessment is presented in Section 4; in Section 5, we conclude and outline our future research plans.

DATA COLLECTION

Our analysis is based on the data set collected from LiveJournal—a popular massive online blogging social network site (BSN). A BSN allows individual bloggers to form contact lists, subscribe to their friends’ blogs, comment on selected blog posts, declare interests, and participate in communities—collective blogs. Thus, a blogging network is a bimodal venue where users engage in both publishing and social activities [13]. As of Spring 2012, LiveJournal has 32 mln individual and community accounts. A LiveJournal user maintains his/her personal blog (public or private) and...
may be a member of an unlimited number of special- and general-interest communities.

We identified 43 NSSI-related communities in LiveJournal\(^1\). Some of these communities promote NSSI activities, while others advocate for NSSI abstinence. We collected the self-declared interests of 22,000 LiveJournal users who are members of or participate in the selected communities (the total of \(\sim 150,000\) interests, including misspelled, abbreviated or hyphenated variants). Participation is defined as posting to the community blog or leaving comments to community blog posts or to other comments made by other participants.

Thus, we formed a matrix \(M\) where \(M_{ij} = 1\) iff the user \(U_i\) has declared the interest \(V_j\), and \(M_{ij} = 0\), otherwise. In other words, \(M\) is the adjacency matrix of a two-mode network of users and their interests.

**SEMANTIC NETWORK GENERATION**

We use the matrix \(M\) to generate a semantic network [11] of interests corresponding to the NSSI population. This network is a one-mode projection of the original two-mode network induced by the matrix \(M\). It is undirected, weighted, and signed. The nodes of the network represent interests \(I_i\) and the edges represent the corresponding general similarities \(C_{ij} \in [-1, 1]\).

Thematic (e.g., NSSI-related) communities are more homogeneous than general-interest communities. They consist of people who are similar in a certain sense. In an extreme case, all community members would be uniformly interested in the community subject and use common terminology. This similarity should be taken into consideration while calculating correlations between declared interests. It has been shown by Kovacs [5] and confirmed by our finding that agent agnostic Pearson correlation underestimates the proximity of terms. Kovacs generalized similarity measures take the population structure into account. They are defined recursively: two terms are similar with correlation \(\Theta_{ij}\) if they are used by similar people; two people are similar with correlation \(\Phi_{ij}\) if they use similar terms (\(\Theta_{ij}, \Phi_{ij} \in [-1, 1]\)).

Let \(M_i = M_i - \overline{M_i}\) and \(M_j = M_j - \overline{M_j}\) be the \(i\)th row or the \(j\)th column of the matrix \(M\), respectfully, centered by subtracting the mean of the corresponding row or column. Then matrices \(\Theta\) and \(\Phi\) can be calculated recursively by starting with two appropriately sized identity matrices \(I\):

\[
\Theta_0 = I, \quad \Phi_0 = I,
\]

\[
\Theta_{i,j,k+1} = M_i \Phi_k M_j^T / \sqrt{(M_i \Phi_k M_i^T) (M_j \Phi_k M_j^T)},
\]

\[
\Phi_{i,j,k+1} = M_i^T \Theta_k M_j / \sqrt{(M_i^T \Theta_k M_i) (M_j^T \Theta_k M_j)}.
\]

After a number of iterations the algorithm converges to the "true" values of \(\Phi \approx \Phi_\infty\) and \(\Theta \approx \Theta_\infty\). The similarities between community members \(\Phi\), though calculated, are not used in this study.

\(^1\)A complete list of communities with their posting and membership statistics, etc. is available from the authors in electronic form by email.

As a side note, in the case of totally heterogeneous population, \(\Theta = \Sigma\) and \(\Phi = 1\) (each person is similar only to herself).

By construction, \(\Theta\) is a dense symmetric signed square matrix with few or no zero terms. The distribution of similarity measures in the matrix is close to uniform.

Calculating \(\Theta\) for 150,000 interests is computationally infeasible due to time constraints and arithmetic imprecision. We restricted our study to the top 600 most often declared interests shared by \(\sim 14,000\) NSSI bloggers. That was the largest matrix that could be evaluated on a 64-bit AMD desktop computer with 8GB of RAM.

**SEMANTIC NETWORK ANALYSIS**

To explore the organization of the semantic network of interests, we extracted some of the strongest generalized similarities between the interests by creating another adjacency matrix \(\Psi\):

\[
\Psi_{ij} = \begin{cases} 
\Theta_{ij} & \text{if } \Theta_{ij} \geq 0.8 \\
0 & \text{else}
\end{cases}
\]  

(1)

Matrix \(\Psi\) is square, sparse (its density is 12%), symmetric, undirected, weighted (in a limited range), and unsigned. It has 42,000 non-zero entries that correspond to 21,000 network edges.

We used program Gephi [1] to visualize the network described by matrix \(\Psi\). The sketch of the network is shown in Figure 1\(^2\). The network has a clear hierarchical structure.

\(^2\)The detailed network map is available from the authors in electronic form by email.
It consists of four major clusters: “music” (MUS), “pathology” (PAT), “daily life/emotions” (DLE), and “creativity” (CRE). Some most frequently declared interests from each of the clusters are shown below:

**MUS:** atreyu, him, incubus, korn, my chemical romance, nirvana, rancid, system of a down, the perfect circle;

**PAT:** alcohol, anorexia, bulimia, burning, cutting, handcuffs, pain, self-injury and self-mutilation (both with and without the dash), spikes, weeds;

**DLE:** cameras, cloths, dvds, flirting, flowers, fun, quotes, smiling, hearts (also as an HTML entity &hearts; and as ♥);

**CRE:** astrology, books, languages, philosophy, psychology, shakespeare, sociology, travel, wine.

There is surprisingly little connectivity between the clusters CRE and MUS. The remaining border zones are spanned with few important bridge interests:

**PAT/MUS:** (black) eyeliner, girl interrupted, metal;

**PAT/DLE:** candy, girls, insomnia, red, rock music, sex;

**MUS/DLE:** animals, camping, fashion, games, honesty, humor, travel(l)ing;

**All four clusters:** bands, bracelets, hoodies, lesbians, making out.

**SEMANTIC NETWORK COMPARATIVE ASSESSMENT**

While many of the associations shown in the map in Figure 1 may be specific to the NSSI community members, some may be either totally random or specific to the age or cultural group to which these members belong.

Thus, some associations that seemingly suggest NSSI behavior, may turn out to be misleading. For comparison, we used the technique described above to build interest maps for some non-pathological communities with similar demographics.

LiveJournal users can post their date of birth to the profile page, but not gender. This, and also the fact that on average only 25% of users report their age, makes it challenging to compare the demographics of communities. After some research we identified two communities that reasonably fit our criteria: “sexy-mood-music” (SMM, 6,700 members, average member age 25 years—same as in the NSSI communities) and “movies-in-fifteen-minutes” (M15M, 13,300 members, average member age 28 years). The communities are catering to music and video fans, respectively.

We collected the top 450 and 550 most frequently used interests of the SMM and M15M members and generated their semantic maps \( \Psi_{SMM} \) and \( \Psi_{M15M} \). We then calculated the intersection between the NSSI semantic network and each of the other semantic networks under consideration. The intersection contains the associations that are significant for both communities and presumably are pathology-free.

![Figure 2. Two intersection networks showing common interests NSSI ∩ SMM and NSSI ∩ M15M](image)

We combine semantic network edges using fuzzy set theoretical operations for intersection and difference:

\[
A \cap B = \min (\alpha, \beta) \\
A \setminus B = \min (\alpha, 1 - \beta)
\]

Here, \( A \) and \( B \) are edges, and \( \alpha \) and \( \beta \) are generalized similarities associated with edges.

Let \( \Psi \cap x \) be the intersection of the original network and a comparison network \( x \in \{SMM, M15M\} \). Then for each edge \( e_{ij} = (V_i, V_j) \) in \( \Psi \), if the corresponding edge also exists in \( x \), then this edge is added to the intersection network with the weight calculated using Eq. 2. Otherwise, its weight is taken to be 0 to emphasize the lack of either similarity between the two terms:

\[
(\Psi \cap x)_{ij} = \begin{cases} 
\Psi_{ij} \cap x_{ij} & \text{if } e_{ij} \in (\Psi \cap x) \\
0 & \text{else}
\end{cases}
\]

The resulting intersection networks for \( x \in \{SMM, M15M\} \) have fewer nodes (300 and 220) and about the same density (9% and 12%). We analyzed them using Gephi software (Figure 2) and discovered that they have a remarkably common structure: both networks have dense and strongly connected CRE and DLE clusters and smaller and less connected to the “mainland” music/movies clusters. The latter clusters, in turn, consist of easily identifiable “movies” (MOV) and MUS subcomponents.

This structural similarity suggests that the intersection networks represent some common base that reflects typical adolescent development [6].

Next we adjust the original network with respect to the comparison networks as a way to better distinguish the features of the NSSI semantic network.

Let \( \Psi \setminus x \) be the difference between the original network \( \Psi \) and a comparison network \( x \). If edge \( e_{ij} \) exists in \( \Psi \) but not in \( x \), then it is essential and is inserted in the difference net-

3
work with its original weight. If the edge exists in both networks, it is inserted with the weight calculated using Eq. 3. Otherwise, the edge exists only in $x$; it is irrelevant and is not inserted:

$$(\Psi \setminus x)_{ij} = \begin{cases} \Psi_{ij} \setminus x_{ij} & \text{if } e_{ij} \in \Psi \\ 0 & \text{else} \end{cases}$$

The adjusted NSSI interest network with respect to $x = M15M$ is shown in Figure 3. The new network has much less dense CRE and DLE clusters; cf. the original network in Figure 1. The PAT and especially MUS clusters are still very dense. From Figure 3, we can identify two groups of possible NSSI beacon interests:

**Non-PAT interests in PAT:** angelina jolie, bdsm, beer, being alone, bisexuality, black, boots, crying, dying, fire, fishnets, goth(ic), graveyard, hair dye, horror, industrial, lust, perfection, porn, serial killers, sex, tattoos, tears, vampires, wicca, witchcraft, etc.

**Bridge interests:** anxiety, bracelets, corsets, edgar allen poe, emotions, (black) eyeliner, girl interrupted, girls, glitter, horror movies, insomnia, leather, lesbians, magick (sic), marylin monroe, (heavy) metal, night, poems, red, safety pins, screaming, spikes, techo, tori amos, etc.

Our findings also appear indicative of the growing global middle-class youth culture revolving around leisure activities (e.g., music, art) reflecting adolescent development in internationally-connected networks [6]. This is further supported by the similarities between the NSSI interested communities and the non-pathological comparison communities. Notably both sets of communities included entertainment, creativity, and daily life clusters.

**CONCLUSION AND FUTURE WORK**

We constructed a semantic network of interests declared by non-suicidal self-injury (NSSI) bloggers of LiveJournal. The network consists of four clearly separated interest clusters corresponding to the pathological terms (e.g. “self-injury” and “razor”), daily life, popular music, and creativity. The interests that bridge gaps between the pathology cluster and the other three clusters can be used as beacons signaling the potential presence of an NSSI behavior.

In related research individuals with a history of NSSI are found to view themselves negatively (e.g., less intelligent and more emotionally unstable) and as having lower social capital (e.g., less attractive, weak social skills [2]). The extent of MOSN NSSI-related communities on LiveJournal could evidence the limited opportunities for social networking among people (e.g., self-harmers) who find themselves excluded from their local communities/local peer networks. Future work is needed examine how members of NSSI-related communities use MOSNs to affirm a sense of meaning and obtain social support and expanding social capital. At the same time, increased time in unstructured peer interactions via NSSI-related MOSNs may lead to further involvement in deviant and antisocial behavior in early adulthood [6].

In this study, we considered only self-declared interests displayed on the bloggers’ profile pages. Some of these interests may have been chosen randomly or based on certain (sub)cultural considerations, and do not necessarily reflect the real user’s attractions. As the next step in this direction, we plan to study keywords in the messages posted to the NSSI communities. We expect that the free-form language of the messages is a better proxy for the pathological behavior. If our hypothesis is right, then the semantic network generated from the keywords will differ from the network $\Psi_-$ constructed in this study. The area of overlap will probably be the cluster of “true” NSSI interests.

**Acknowledgment**

This research has been supported in part by the College of Arts and Sciences, Suffolk University, through an undergraduate research assistantship grant. The authors are grateful to Zoë Wells of Suffolk University for preliminary data collection and Dr. Jim Hollander and Prof. John Boyd for suggestions on combining graphs.

**REFERENCES**


among adolescents: a critical review of the literature. 


5. Kovaćs, B. A generalized model of relational similarity. *Social Networks* 32, 3 (July 2010), 197–211.


