Information Revelation in Online Social Networks

Jernej Zorko
Faculty of Electrical Engineering and Computer Science
University in Maribor, Slovenia
Mathematics and Engineering Department
University of Madeira, Portugal
jernejz5@gmail.com

ABSTRACT
In this paper I will present my study on a Slovenian news portal Vest.si. The goal of my study was to understand how and which personal information is revealed through posting in this online social network. Because posting in this online community can either be anonymous or not I have studied which kind of personal information contributors when being anonymous fell freer to reveal in their posts. To get data I used empirical study of this online community during a period of one month in fall 2009.

Authors Keywords
Information Revelation, Vest.si, posting, commenting empirical research.

1 INTRODUCTION
Online social communities have become a part of everyday routine in the recent years for a large number of people. We are using them to communicate with friends, to understand them, to know what they are doing or thinking, to express ourselves and to get and share different information or goods. To be more precise let us look at Facebook1 which is one of most popular social sites in the last year. Their instant messaging function is connecting people and has 1 billion messages sent per day via Facebook chat. That is 4-5 chat messages per person per day in average. (1) If we look further into Facebook statistic page2 we can see that 50% of active users access their account daily and that every active user updates his status in average 0.15 times per day, and shares 1.5 pieces of content per day. Facebook Data Team blogged a research where he focused on content of Facebook status updates. (2) Status updates in Facebook can either be studied at a matter of public trends or from more private perspective which becomes interesting because of four properties of mediated publics. Those are: Persistence, Searchability, Replicability and Invisible audience. (3) This means that information we leave in social community as Facebook will stay published, users will be able to search it and we will not know for sure who read it or not. In my research I chose online social community that offers users to decide if they want to participate anonymously or not. I will study information that is being revealed through participation in this community.

1.1 VEST.si
Main goal of vest.si3 website is to provide visitors with current news from Slovenia and abroad. Vest.si started publishing articles in May 2007. Direct translation of word vest from Slovenian can be interpreted as both conscious and as single news. Their way of providing news to public is a bit different from other Slovenian sources. Fourteen starters of this portal wrote their presentation when the portal was published for the first time: “we are the riders of new technologies, the critics of over lived patterns and messengers of news. Good news do not need any outfit, just a path to the ears. Question of yellow print is unreal, there is just decency. Argument is the king; right to correction is sacred, critical judgment preserves common sense. If a story is being watched, it will be revealed by her own.” It also states as towards the reader: “You forgot: News is not to be traded, the truth is not an idea. Truth is one, given to no one. Contempt will be replied with contempt. You were a servant, and you will stay one.” Then there are given seven guidelines for the journalists to respect them and guide their articles: “You, who are being called a journalist: Respect single news and every person. Report clearly, exact, and in well augmented language. People have names, places live in their time. Take notes of public affairs, respect private ones. Choose carefully your style and genre of writing. Be without mercy when revealing unrespectable acts. Without the source, a story is just a story. Feel your conscious.” Those statements are written and signed on vest.si presentation site.4

I think upper paragraph offers reader of this paper a good insight of how creators of www.vest.si are preparing and publishing content on the internet. Most of the creators when they created this portal were already either well read bloggers in Slovenia, promising young journalists or middle aged journalists that for different reasons were not very known writers. Combination of their youth, different

1 www.facebook.com
3 www.vest.si
4 www.vest.si/o-vesti
aspects of their lives and different experience levels, created great team that searches and prepares news on their own successful way so a lot of people in Slovenia read their contents daily.

• **Politics**
  Reports and comments of current political affairs mostly from Slovenia, but also from Europe and World.

• **Society**
  Comments on most known social events and social life in Slovenian area.

• **Culture**
  Articles about cultural events in progress. This group often serves as a guide to some interesting events that do not get much media support.

• **Fun**
  Usually visitors find videos that trough different aspects of our everyday life often offer critical view on society

• **Opinions**
  Video articles of guests that present their versatile, often provocative opinions.

• **Projects**
  Variety of projects that are being differently connected or sponsored by vest.si

• **Columns**
  57 different public columnists write their articles that are usually one of the most commented content of vest.si.

• **Blogs**
  Public users sharing their opinions. For now there are active 30 bloggers.

All video material, also video inserts in articles, is published on YouTube⁵ or Vimeo⁶

Users contribute on vest.si online social community when they are posting comments. They can post comments on every article, blog, show or column and also on other comments. If a user decides to comment some content he has to fill in his name, e-mail, text and URL. URL textbox is usually used to prove your identity. Users fill in their blog URL usually. In this kind of a system it is not necessary to fill in your true name. Users can use aliases or even every time different name. But this usually does not happen because users want to backup their opinions with their social status which is hidden behind their alias or name. In my research I have also come across user stealing else identity. One of the users used alias of some other user and wrote a comment that was shortly recognized not to be from the right person because content of the comment did not match that user’s behavior. User was usually criticizing work of one of Slovenian politician and in that case he was doing the opposite, and when one of the observers replied his

---

⁵ vimeo.com/omg
⁶ www.youtube.com/user/vestinavesti
assumption also owner of the alias confirmed what happened.

1.2 Other Studies

Ralph Gross and Alessandro Acquisti studied information revelation on Facebook on CMU students. They studied which information Facebook users fill in their personal profile and found that most than 50% of user keep their address, phone number and summer job field private. Less than 20% of users keep their profile image, birthday and high school field private. Other fields are between those values. They also found that 89% of users use their real name, 61% of users use profile picture that is suitable for direct identification (this percentage is dramatically different compared to Friendster). Combining all the data they came up to some interesting result: 16% of female and 21% of male user’s identity and location can be discovered. (3) If we would ask users if they feel like their identity and location can be discovered and if they are worried about it I think we would be surprised by the results. A study of OUT-LAW tested how hard it is to get Facebook friendship of an unknown person. According to their research 46% of young people accepted friend request and revealed personal information to strangers. 10 years ago we would need a con-artist to get this kind of information about a person. (5) Harvey and Soltran study can also serve some interesting results regarding privacy information revelation. Half of studied users are not or are a just a little concerned about their privacy. There are also some gender differences: men talk less about themselves. (6)

1.3 My Focus

In my research I have focused on web site Vest.si with purpose to study how is personal information being revealed trough principle of commenting content. The Executive Office of the President of the United States of America defined personal information as information which can be used to distinguish or trace an individual’s identity, such as their name, social security number, biometric records, etc. alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother’s maiden name, etc. (8) In European Union a similar term personal data. It is defined as any information relating to an identified r identifiable natural person; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity. (7) In Australian law they have a little more technical definition used also for Health Records. They define personal information as information or an opinion (including information or an opinion forming part of a database and whether or not recorded in a material form) about an individual whose identity is apparent or can reasonably be ascertained from the information or opinion. (9) (10)
The purpose of my study is to find any parts of personal information in online users participation records. I am arguing that users protected by their online anonymity are more likely to reveal personal information. I think that users feel freer if their identity is not known. To confirm that I had to find which information are users more likely to reveal, how often does that happen and how users use their privacy options.

2 METHODOLOGY
Main goal of this research was to observe social behaviour on my portal and study it. After defining my objectives and focus I decided to use empirical (observation) method. Next step was collecting data from website. In the beginning I was hoping on cooperation with site owners, but after some requests sent to Vest.si administration they all hit dead end. Then I focused on different methods saving comments data. After some observation I have found that the easiest method to retrieve comments data is via Vest.si comment RSS feed. To manage the data I used Google reader, web application that helps user to save, view, mark the RSS elements and supplies us with basic statistical analysis. In the beginning I was expecting up to 50 comments posted per day. That is the reason I started thinking of getting a secondary source of data. After some research I came up with option of using web crawling on my target website. I found a directly enumerated link to all articles in the URL www.vest.si/?p=30150. P parameter in that case represents article identification. Using .NET platform I started working on a desktop application that retrieved html data from URL, from different range of parameter p in the URL. When testing the Vest.si server permissions I retrieved 1080 html pages in 3 hours. Interval parameter was set to 10 seconds. Purpose of second stage of establishing crawler was to retrieve comments data from html code using .NET XML handlers. After successfully realized second stage already in average 140 comments per day were retrieved via feed. That in reality takes up to half an hour to read every day. Time depends of topic. My plan was to read comments for one month, so by simple calculation this means around 4000 comments. That was by my estimations enough for my study. I was expecting in average for every 20th comment to be revealing personal information. That would in the end bring 200 personal information revealing comments.

After collecting data analysis took part. In process of analyzing individual comment first was reading it. If comment included some personal information it was bookmarked as “comment useful for my research” and prepared for continuing. After collecting useful comments, they were tagged by type of information. Those types are:

- Education (Parts or full information about ones educations. This Includes place of study, facility and course)
- Age (Revealing information on participant’s age)
- Location (Different information about locations of living, working or maybe just revealing of current location)
- Work (Information about profession.)
- Identity (Full name or just parts of users name combined with some other types)
- Relations (Relations information are usually based on referring to people user knows. Sometimes users want to back up their statements by referring to family, friends, siblings or maybe some people that they shared experiences)

The last part of analysis was determining the level of information revealed by categories:

- Low (If occurrence is marked as low. This means that information cannot be identifying but in case of connecting it with some other information this could be possible)
- Medium (Medium category means that information is partly identifying)
- High (High category is directly identifying)

3 RESULTS
My data collection lasted from 13th of November 2009 to 13th of December 2009. After that observation my collection included 5346 comments data that came from 94 recorded articles. In average that numbers mean 3 articles per day, 172 comments per day and 57 comments per published article. Comment distribution during weekdays is presented in Figure 3.

![Figure 3: Weekly distribution of comments](image)

Out of 5346 comments there were 47 comments useful for my study. That is 0.9 % of all comments. This was a lot under my expectations.
After assigning types to useful comments I have come up with results presented in Figure 4. Most users gave away information on their location and about education. Least users gave away their information on their personal relations.

![Figure 4: Useful comments types distribution](image)

Next stage was determining the level of information revealed. Fixture 5 shows us distribution on education information revealing. Most of the occurrences were not well defined. 9 users defined facility or university where they studied or went to school. Only 3 users defined which course they attended at university. There are no users that would completely define when and what did they study.

![Figure 5: Revealing level distribution on education type of information](image)

Focusing on age, 1 user wrote his age numerically and 1 wrote his year of birth, 2 users wrote indirectly how old are they. For example, one user wrote that when Slovenia became independent he was serving army which means he was from 20 to 25 years old. 1 user wrote he is young. Graphical representation is shown in Fixture 6.

![Figure 6: Revealing level distribution on age type of information](image)

In total 6 contributors wrote about their work. 4 users defined where they work and 2 users defined what their position in company they work. Graphical representation is shown in Fixture 7.

![Figure 7: Revealing level distribution on work type of information](image)

There were 7 contributors that wrote about their identity. 2 of them completely defined who they are. One of those two examples was a female posting complete letter she wrote with her name and address. 3 of them wrote their name and street they live and 2 of them wrote their name and age. Numbers are presented inFixture 8.

Most occurrences of personal information revealances were noted in category location. 2 of the users wrote in which street they live, the most – 12 users wrote name of the city they live in and 2 merely defined their region of working. Fixture 9 represents that.

Last category to study is personal relations. 2 contributors were referring on their previous co-worker and one was referring on personally knowing one of Slovenian politicians. This is presented in Fixture 10.
The last part of my research was to note which of useful comments were anonymous and which were signed. For a comment to be signed, it had to fulfil criteria: signed with name used more than once, either this can be real name or alias which can be connected to real name (some contributors use the same aliases that they use on their blogs). If I was able to get their identity information I marked comment as signed. Otherwise it was marked by unsigned. Results by category show that in category education 3 high level comments were signed. In category age 1 medium level comment was signed. In category location 2 high level and 2 medium level comments were signed, meanwhile in category work 1 high level and 1 medium level was signed. In category identity 3 high level and all three medium level comments were signed. All comments in category relations were unsigned. This in total means that 30 % of comments were signed. Fixture 11 shows that values in partial presentations.

The CONCLUSIONS AND DISCUSSION
I was surprised over number of useful comments. I was expecting number around 5 % of all comments to be revealing information but on the end of my research the number was much lower. Only 0.9 %. I find commenting distribution interesting. I think it shows when contributors have the most time to use online news sources. Sunday and Monday came out as least popular for contributing.

After grouping comments I confirmed my thinking that most users will reveal their location and education. I think this is because Vest.si has mostly political issues and some statements are stronger if supported by strong identity of user.

Comments grouped by type of revealing show that contributors feel freer to give away more detailed information about their work and age and the least detailed about their education and location.

A bit surprising were the numbers on signing comments. I was expecting bigger percentage of signed comments at education and relation however results show that users feel most free to sign their statement about their work, identity and education. Also surprising is that no comment in group relation was signed. I am explaining this so that people who back up their statements using relation references are usually newbie’s or not so experienced members and do not feel comfortable to sign their statements.

I was arguing that users protected by their online anonymity are more likely to reveal personal information. By my research this statement was confirmed. 30% of comments that revealed some personal information in part or whole was not signed in the way that signature can be interpreted into real life identity.
I will also like to add that information gained from vest.si may not be entirely correct. One of possible reasons is short observation time. I think research would be a lot more representative if it would be possible to collect and analyze data over longer period of time. I think that if I would have that amount of data and analyzing compatibility cross reference of data may also bring a lot of interesting results. I think that if we would do that by aliases we could combine all information about one person. I am wondering how much this would be. Also Vest.si’s topics are quite sensitive for discuss. Those are political views and I think some people do not feel comfortable signing those statements. One of possible reasons is fear from that those information could harm them some time later.

5 BIBLIOGRAPHY


6. **Facebook: Threats to Privacy.** Jones, Harvey and Soltren, José Hiram. 2005.


