"1st International Symposium on Computing in Informatics and Mathematics (ISCIM 2011)"
in Collabaration between EPOKA University and "Aleksandër Moisiu" University of Durrës
on June 2-4 2011, Tirana-Durres, ALBANIA

Analyzing the Virus Spread over Facebook Social Network using Descriptive Data Mining Techniques

Valentina TOMOSKA [(1)], Özcan ASİLKAN [(2)], Katerina RISTOSKA [(3)]

(1,3) SEE University, Faculty of Contemporary Sciences and Technologies, Tetovo, MACEDONIA;
Email: vt11573@seeu.edu.mk, kr11267@seeu.edu.mk
(2) Epoka University, Department of Computer Engineering, Tirana, ALBANIA
Email: oasilkan@epoka.edu.al

ABSTRACT

Social networking currently has become one of the most popular forms of communication among users on the Internet. Accordingly, with the exponential growth of users participating in the social networks and their exposure of sensitive personal information online has added additional value to the security breach issue. Namely, by curiously accessing various fraudulent links or applications that social networking sites offer to their users has provided a "Petri plate" for rapid spread of infectious viruses. To make things even more complicated, the friend-of-a-friend structure of these types of networks is correlated both with the speed of spreading, as well as the direction of spreading of the viruses. Therefore, this instigated the conduction of the survey on how viruses spread across the Facebook network, out of which some descriptive data mining statistics were gained and analyzed in the IBM's SPSS Statistics tool, along with a comparison of these results between the examined participating countries (Macedonia, Albania and Kosovo).

I. INTRODUCTION

Facebook, a social network that facilitates people in interconnecting with their friends, family and colleagues, has been founded in February 2004. It contains one of the largest MySQL database clusters, which according to their statistics has more than 500 million active users, where an average user has 130 friends, is connected to 80 community pages, groups and events and creates 90 pieces of content (web links, news stories, blog posts, notes, photo albums, etc.) each month (or an average of 3 pieces of content per day)[21]. Hence, briefly and informally stated, social networking has become the root through which information is shared on the Internet. However this is not limited to sharing only useful information; in

[21] http://www.facebook.com/press/info.php?statistics

most cases, the data flow involves some hidden content which is being distributed around the social network users' profiles, as it is the case with many Facebook viruses circulating through countless links, applications or similar [3,4]. Specifically, Facebook viruses penetrate and hijack sensitive account information by spreading via the above-mentioned pieces of Facebook content types, as in the case of "Koobface" malicious worm (whose name is an anagram of the term "Facebook"), which breaches through Facebook messages and by its fast replication, causes network traffic congestions [2,5,6]. Thus, keeping this in mind, one might begin to wonder how to improve the privacy settings regarding information sharing on a site like Facebook; consequently, this was the reason which inspired this study for researching how these viruses might spread across the Facebook network Research employed a public survey on Facebook users (participants) from three countries: Macedonia, Albania and Kosovo. Collected data were later examined more thoroughly through some data mining descriptive statistics.

Data mining as a concept is the extraction of hidden predictive information from large datasets used to analyze and predict future tendencies and behaviors that enable preemptive knowledge-driven choices in organizations [1]. Data mining offers fast data processing through tasks like: interactive and visual simple exploratory data analysis, descriptive modeling which offers models for probability distribution, p-dimensional space partitioning and variable dependence models, predictive modeling which acquires from previous patterns or variables and predicts future outcomes, discovering patterns and rules of falsified data and data retrieval by content (pattern), mostly used for text and image datasets. The data mining models are divided into **predictive** (classification, regression, time series analysis, prediction and other), which make prediction about unfamiliar data values by using the identified values and **descriptive**, which detect the patterns or relationships in data and discover the properties of the data observed (clustering, summarization, association rule, sequence discovery or similar) [7]. These models inspect the data through data cleaning, integration, selection, transformation, extraction, evaluation and visual representation of the gained knowledge by using statistical algorithms, decision trees, nearest neighbor algorithms, neural networks, genetic algorithms, rule-based algorithms, support vector machines etc [1]. All of these mentioned tasks and algorithms are currently implemented in various commercial data mining tools like DBMiner, IBM SPSS Modeler (former SPSS Clementine), DB2 and Intelligent Miner, and open source tools like RapidMiner, Weka, Keel etc. Other such products are the decision tree based CART, Scenario, See5, S-Plus; the rule-induction based WizWhy, DataMind, DMSK; the neural network based NeuroShell2, PcOLPARS, PRW; the polynomial network based ModelQuest Expert, Gnosis, KnowledgeMiner; the association/classification and text mining based TheMining, EPRules, Simulog, Sequential Mining, O3R, KAON, MultiStar, CIECoF; the statistics and visualization based Synergo/ColAT, GISMO, TADAEd (Lile and Asilkan, 2011).

II. RESEARCH

Analyzing Frequencies

Frequency analyses involve describing discrete categories of data having multiple choices in the answers. These analyses occupy constructing a frequency distribution, which is a record of the number of scores that fall within each response category. The frequency distribution includes two elements: (1) the categories of response, and (2) the frequency with which respondents are identified with each category. Frequencies options include a table showing counts and percentages, statistics including percentile values, central tendency, dispersion and distribution, and charts including bar charts and histograms. Thus, these analyses have been used here in order to perform descriptive data mining statistics for the impact on spreading viruses across the Facebook network, that is, in which manner a virus would spread across this social network – via fraudulent links or applications, by analyzing the data frequencies attained in IBM's SPSS Statistics Tool of the previously conducted survey. Namely, the spread of social networks to a greater extent depends on how often the user visits and uses his/her profile. Facebook as a social network is used for different purposes, where some of the users use their profiles just for communicating with their friends abroad and without using any type of applications, links or similar. For Facebook users who visit their profiles more often, a virus has a greater opportunity to spread more rapidly than the others, because of the possibility of using suspicious applications, links and other pieces of Facebook content by the users that access them mainly for entertainment, without foreseeing the type of problems which can be caused due to lack of general network privacy knowledge.

By analyzing how often users visit their Facebook profiles, from the retrieved results it can be concluded that visiting profiles is not too high, only 32.9 percent realize visits very often. The obtained results show that chances are pretty small when it comes to spreading viruses in terms of user's high frequency visits to their profile (see figure 1).

How often do you visit Facebook?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	7	4.1	8.5	8.5
	2	12	7.7	14.6	23.2
	3	17	11.3	20.7	43.9
	4	19	12.3	23.2	67.1
	5	27	17.4	32.9	100.0
	Total	82	53.5	100.0	
Missing	System	74	47.4		
Total		156	100.0		

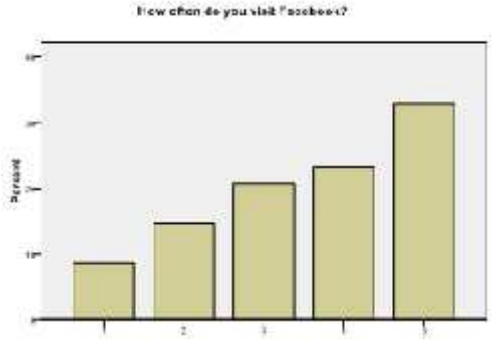


Figure 1: How often do you visit Facebook?

However, if the results of frequently and average visited profiles are considered, the chances of spreading a virus across applications, links or similar, may increase and can cover wider scope. Additionally, the privacy of the Facebook account is very important, because all used applications, links and other objects published on the user's wall can be visible for everyone, causing the possibility of quick spread of the virus. Nevertheless, there are several options for protecting one's Facebook account, as requested by the user, by determining what type of information would be visible and in which proportions to all or just certain group of users. From the results seen in figure 2, it can be considered that in larger percent account information can be viewed only by friends, which is 37.8 percents from the entire examined sample of participants.

Who can view your full Facebook profile?

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	74	47.4	47.4	47.4
Everyone	14	9.0	9.0	56.4
Friends of friends	3	1.9	1.9	58.3
Only friends	59	37.8	37.8	96.2
Only me	6	3.8	3.8	100.0
Total	156	100.0	100.0	

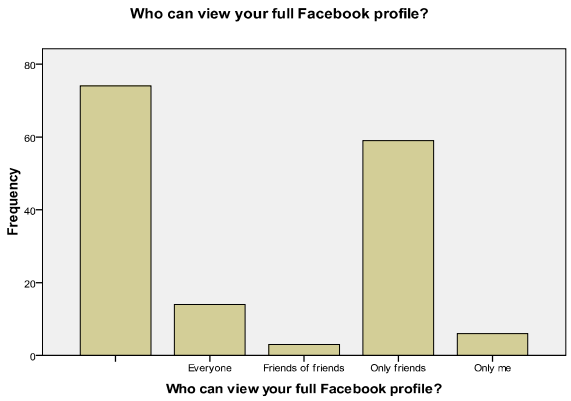
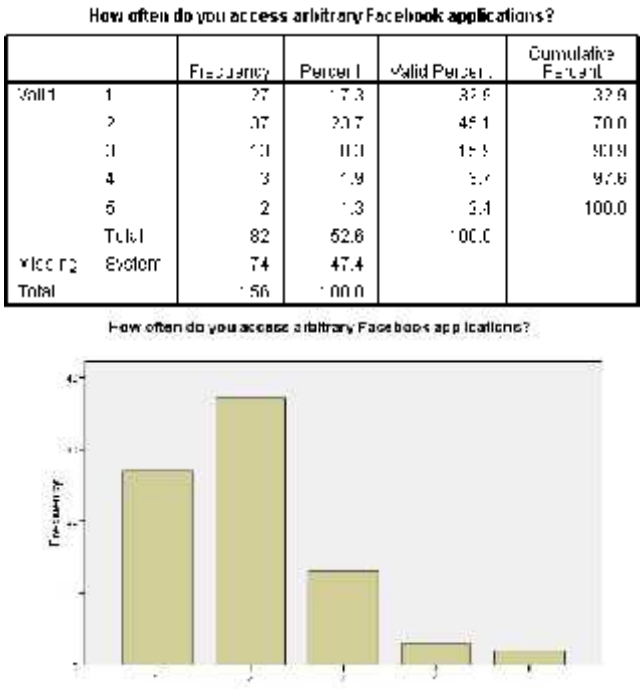


Figure 2: Who can view your full Facebook profile?

The possibility of virus spreading through applications and links is quite high in the close circle of the user's friends and only small percent in the entire network.

The use and accessing Facebook applications by the users is very crucial for the spread of the virus through the applications and the speed of spreading through the network. For wilder spread of the virus through Facebook network or through Facebook accounts, users are offered many interesting and attractive applications, which they access them without much thinking, not assuming that this can cause problems on their computers, infecting them with harm viruses.



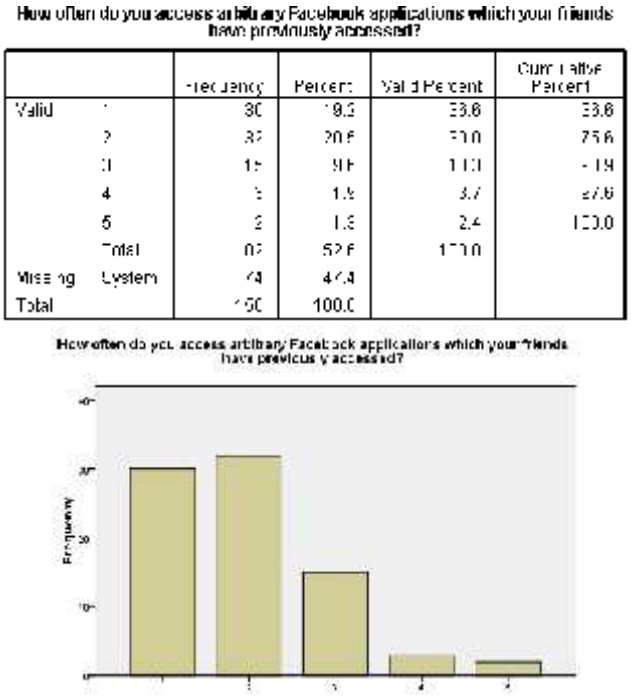
How often do you access arbitrary Facebook applications?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	27	17.3	32.9	32.9
	2	37	23.7	45.1	78.0
	3	11	7.1	13.4	91.5
	4	3	1.9	3.7	95.1
	5	4	2.6	4.9	100.0
	Total	82	52.6	100.0	
Missing	System	74	47.4		
Total		156	100.0		

Figure 3: How often do you access arbitrary Facebook applications?

Considering the results of the analysis (see figure 3), it can be concluded that Facebook users rarely use it for accessing applications and their use, while it is far large percentage of users, 32.9 percents, which rarely used applications or they did not use at all. These results suggest that Facebook users do not access many applications offered by the network, so it can be concluded that the method of spreading the virus through using applications would not give high scores to reach the desired goal.

Often people who are users of social networks are interested in the approaches and actions that their friends make while they are participating in the social network and as well as for those who are not their friends in the network. In this case, interested users start to visit the profiles of their friends (and also profiles of those who are not their friends) and looking at what they did at the moment and in the past, where there is strong possibility if there is interesting and attractive applications used by their friend, the user to access them. At the point when the user wants to visit profile of those who are not their friends, it is very important if the account is set to private or not. So the information of the Facebook account can be locked for everyone, but there are also several options that can be selected and specified for a particular piece of data to be visible to everyone, friends of friends or just to friends. From the obtained results it can be concluded that 19.2 percent do not use applications at all or very rarely, which is very high percentage compared with the percentage of users who use applications, they are only 2.4 percent.



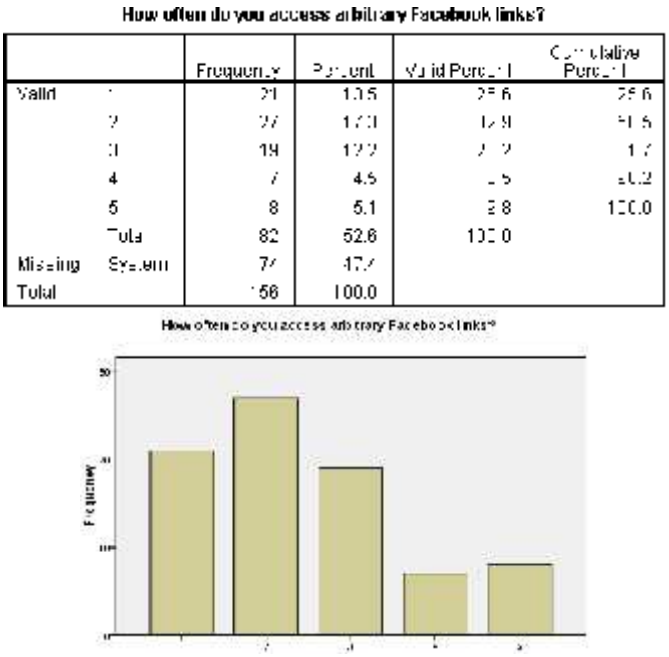
How often do you access arbitrary Facebook applications which your friends have previously accessed?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	80	39.0	39.8	39.8
	2	83	40.5	41.3	81.1
	3	15	7.3	7.5	88.6
	4	5	2.4	2.5	91.0
	5	2	1.0	1.0	92.0
	Total	185	90.2	92.0	
Missing	System	16	7.8		
Total		201	100.0		

Figure 4: How often do you access arbitrary Facebook applications which your friends have previously accessed?

In this part of the analysis, an increase in the average percentage (medium, 3) can be noticed, where previously it was stated that 15.9 percent of the users access applications, while 18.3 percent access applications that their friends have already used. Thus it can be determined that users are accessing application that their users have previously accessed, and in this way the virus is more likely to spread through the network.

Another possible way of spreading the virus in the network is through links, where most of the users don't have the ability and prior knowledge to recognize if the links are infected with a virus or not. From the results it can be seen that Facebook users are not accessing many links, that is, only a percentage of 9.8 of the responders do. The use of the links rather depends on the number of friends the user has, so with increasing the number of friends also increases the possibility for greater use of the links (see figure 5).

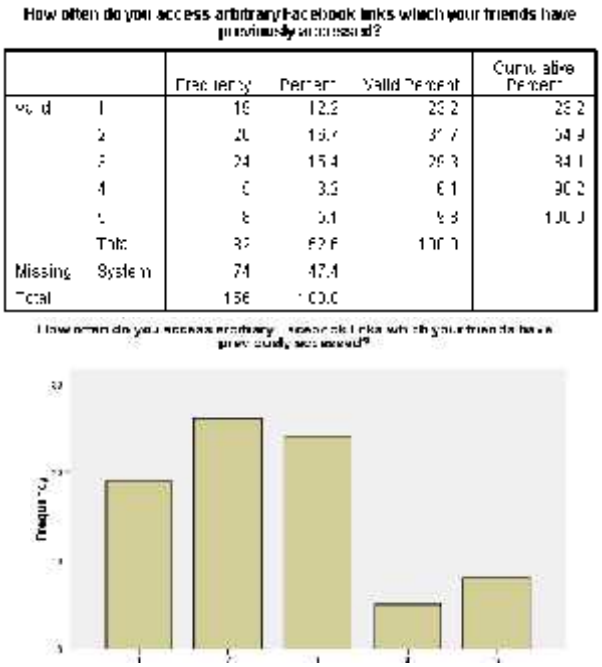


How often do you access arbitrary Facebook links?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	21	13.5	25.6	25.6
	2	27	17.3	32.9	58.5
	3	19	12.2	23.2	81.7
	4	7	4.5	8.5	90.2
	5	8	5.1	9.8	100.0
	Total	82	52.6	100.0	
Missing	System	74	47.4		
Total		156	100.0		

Figure 5: How often do you access arbitrary Facebook links?

Similar to accessing Facebook applications, accessing Facebook links also depends on how often Facebook users use links that were previously accessed by their friends. This is the most successful way of spreading the virus across the network, so that with one access to some infected link, the user unknowingly sends messages to their friends containing the same virus. With large number of friends, the possibility of using infected or uninfected links is greater. Very important and influential factor in this case is the number of friends of the potential user and the speed with which this number increase, since if the number of friends increase, also increases the opportunity for the user to access more potential links that may be infected or not.



How often do you access arbitrary Facebook links which your friends have previously accessed?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	19	12.2	28.2	28.2
	2	26	16.7	37.7	54.9
	3	24	15.4	29.3	84.1
	4	6	3.8	8.1	90.2
	5	8	5.1	9.8	100.0
	Total	82	52.6	100.0	
Missing	System	74	47.4		
Total		156	100.0		

Figure 6: How often do you access arbitrary Facebook links which your friends have previously accessed?

Comparing with the access of applications that were previously used by the user's friends, it can be concluded that access of links that were previously used by the friends have larger percentage. This means that the virus has greater potential to spread through already used links by the user's friends, unlike previously used applications of friends.

The way the virus can spread through the social network may be different, including the manner of spread through Facebook messages containing potential infected link, pictures, or invitations for applications. Examining whether Facebook users use this way of communication, results were obtained from which can be concluded that the percentage is quite small, but compared with percent for using links and applications, this is slightly higher than it. It is also very important if the Facebook user is able to recognize if some unknown link or application on Facebook contains virus. From the results, most of the participants, namely 32.7 percent stated that they are able to recognize an embedded viral link or application, but 19.9 percent cannot.

Before accessing some unknown link, picture or an application on Facebook, would you be able to recognize if any of these contains a virus?

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	31	19.9	19.9	67.3
Yes	51	[illegible]	[illegible]	[illegible]
Total	155	100.0	100.0	

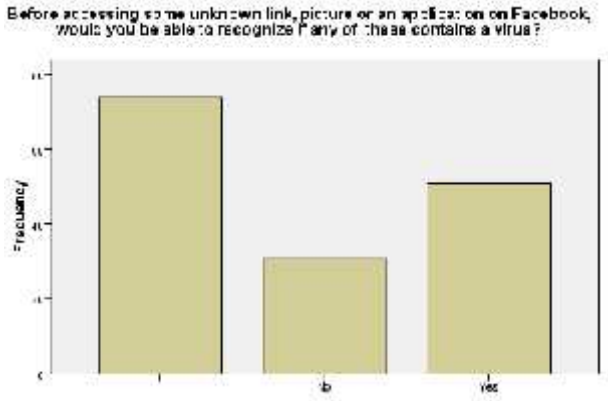


Figure 8: Before accessing some unknown link, picture or an application on Facebook, would you be able to recognize if any of these contains a virus?

This is a very tricky question, because even though users responded positively that they are able to recognize infected links or applications, it is very important that their ability for recognizing is intuitive or recommended by some friends, or more reliably, from a professional standpoint. If the user detects these infected links and applications from a professional point of view, then his answer is more plausible because this user is more likely to identify potential links and applications that contains virus in some other cases.

The survey has covered several regions or countries, specifically, Macedonia, Albania and Kosovo. The most of the participants were from Macedonia (28.8%), then from Albania(19.2%) and a small percentage from Kosovo(2.6%).

Where are you from?

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	77	77.4	77.4	77.4
Albania	19	19.2	19.2	96.6
Kosovo	2	2.0	2.0	98.2
Macedonia	1	1.0	1.0	99.2
Other	1	1.0	1.0	100.0
Total	100	100.0	100.0	

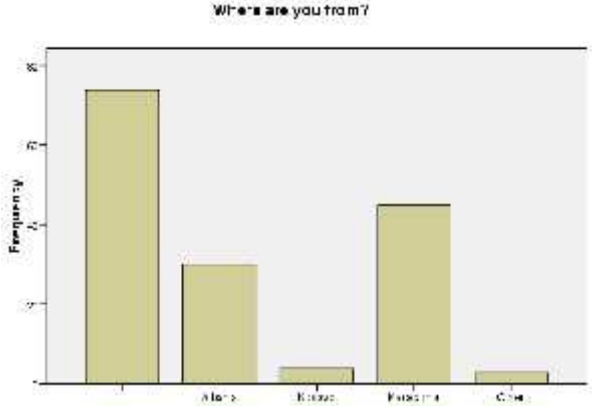


Figure 9: Where are you from?

Analyzing resulting graphs

In this section, the possibility of using the applications and links that are used by friends of the user is being explored. It all depends on several factors, among which the number of friends the Facebook user has, how often the user visits his Facebook account and how often he/she uses any Facebook applications in general. Most often users access applications and links previously used by their friends, who are visiting its Facebook account more frequently and have friends in the range of 500-1000 (figure 10).

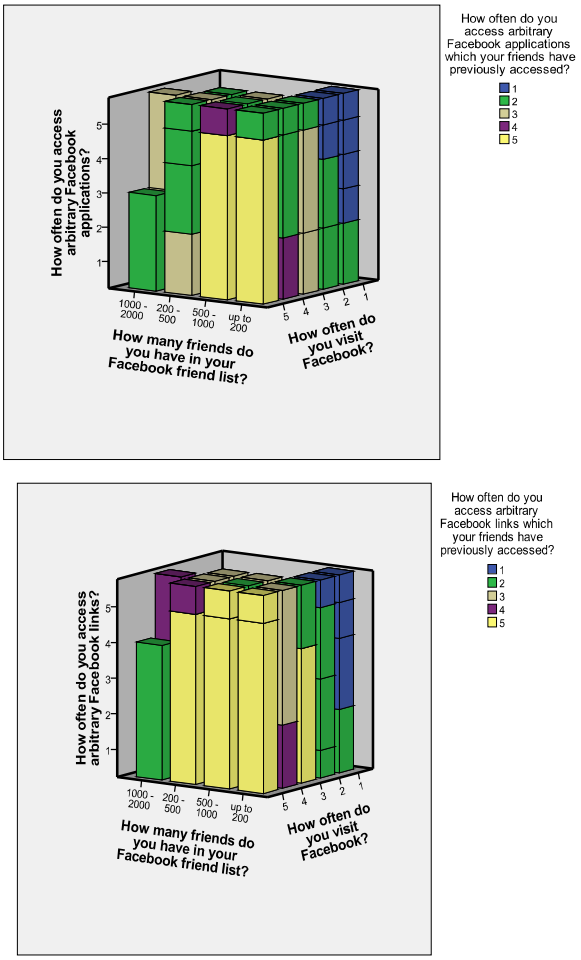


Figure 10: How often do you access arbitrary Facebook applications/links which your friends have previously accessed?

This group of users is quite exposed to viral infections attached to the links and applications, so the spread of the virus is faster and has greater possibility to cover a wider scope of the network.

It has been analyzed how accessing applications are dependent from the privacy of the users account. It is very important and prominent if the user profile is set to private or not. The profile which is not private, all published application and links on the user's wall are available and visible to all Facebook users, and they can access them very often without even being friends of that particular Facebook user (see figure 11).

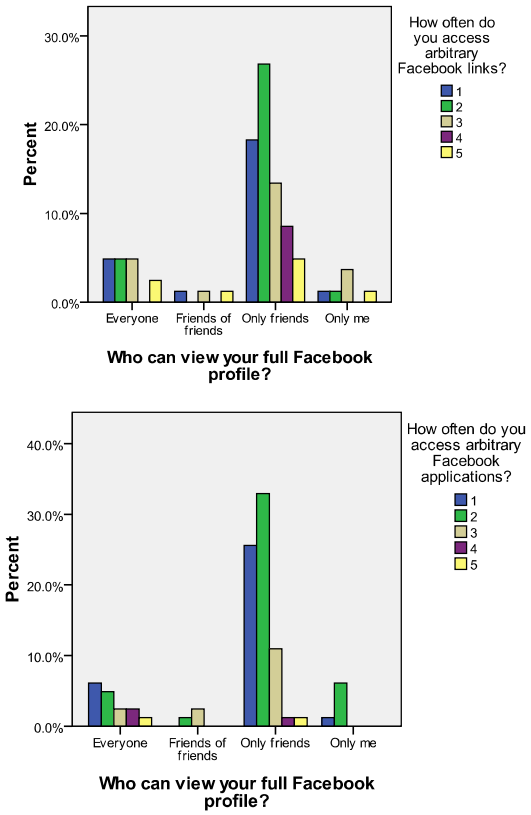


Figure 11: How often do you access arbitrary Facebook links/applications?

From the survey results where the countries involved in the analysis are compared, it can be stated that Facebook users coming from Macedonia have the largest network of friends (see figure 12).

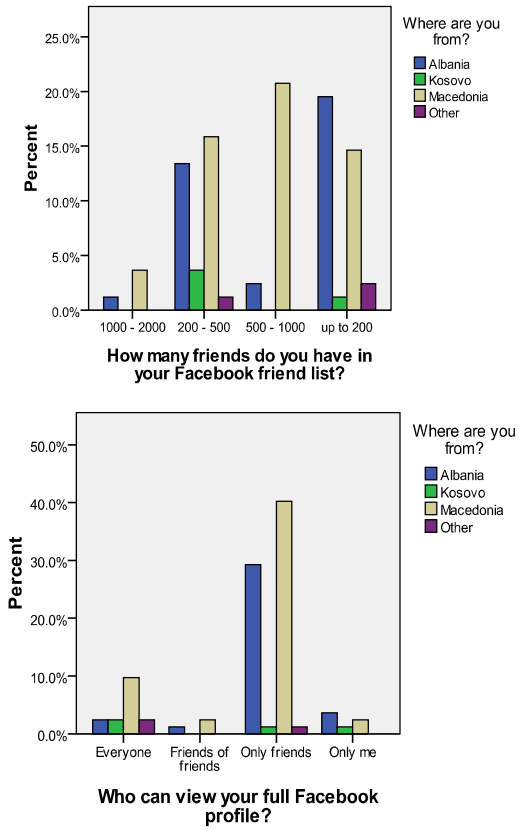


Figure 12: Comparison between the countries in "How many friends the Facebook users have on their friend list" and in "Profile Privacy"

When discussing about the privacy of the Facebook accounts, in all analyzed countries the percentage is highest about profile viewing only by user friends, where the user information is quite protected.

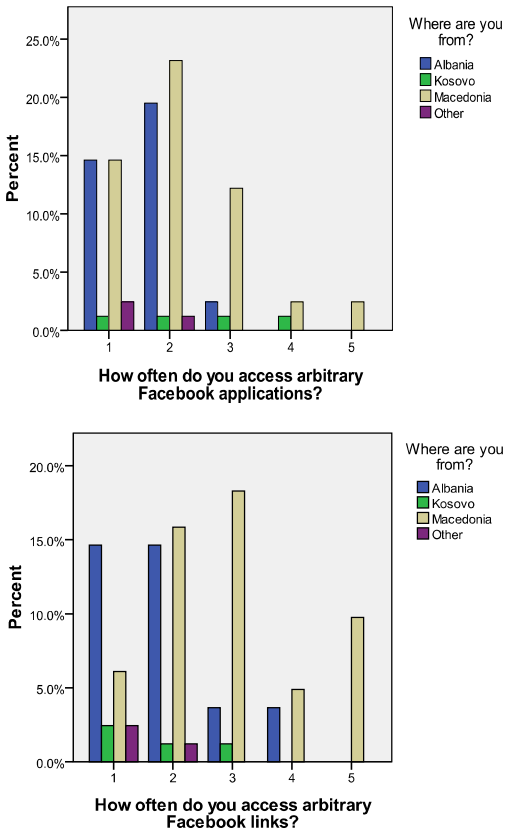


Figure 13: Comparison between the countries in "How often Facebook users access arbitrary Facebook applications" and in "How often Facebook users access arbitrary Facebook links"

The virus can spread faster in the Facebook network for accounts from Macedonia, because those users more often use applications and links, which might be potentially infected by some kind of viruses. Also the Facebook users from Macedonia are more frequently visiting their Facebook profiles, while the users from Albania take the second place, with very low percentage of visits compared with Macedonia.

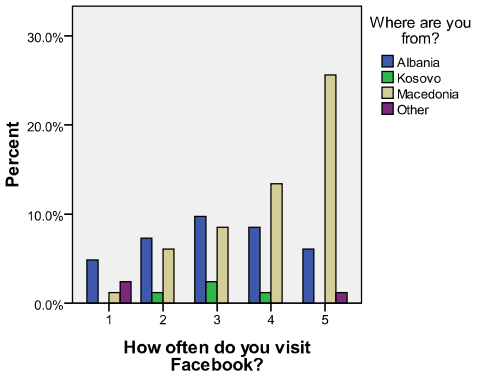


Figure 14: Comparison between the countries in "How often Facebook users visit their profiles".

III. CONCLUSION

This paper has demonstrated that security issues in social networking are rapidly growing with their expansion in terms of participating users. From the conducted survey, which covered three comparing countries, Macedonia, Albania and Kosovo, and used IBM's SPSS Statistics frequency/graph analyses performed on the Facebook links and applications usage, in addition to statistics on the number of Facebook friends, privacy settings etc., it can be concluded that the friend-of-a-friend arrangement of social networks, as well as user's inquisitiveness on accessing different fake links or applications increases the speed of virus spreading across the network and determines its direction. A virus may spread more rapidly via links, rather than through applications and this would eventually occur through Facebook profiles of users coming from Macedonia, than Albania, and the least expected viral spreading occurs at users from Kosovo. Finally, this imposes the necessity for improving the privacy settings on these social networking sites, as well as increased informative broadcastings for educating the social users on how to restrict sharing of personal information online.

IV. REFERENCES

[1] Deshpande, S.P. and Thakare, V.M. (2010) Data Mining System and Applications: A Review. *International Journal of Distributed and Parallel Systems (Ijdps) Vol.1, No.1*

[2] Felt, A. and Evans, D. (2008) Privacy Protection for Social Networking Platforms. *In Proceedings of the Web 2.0 Security and Privacy Workshop (W2SP)*

[3] Goettke, R. and Christiana, J. (2008) Privacy and Online Social Networking Websites.

[4] Govani, T. and Pashley, H. (2005) Student Awareness of the Privacy Implications while Using Facebook. *Unpublished Manuscript*

[5] Gross, R. and Acquisti, A.(2005) Information Revelation and Privacy In Online Social Networks. *Paper presented at the WPES'05*, Alexandria, Virginia, USA.
[6] Jones, H. &Soltren, J. H (.2005) Facebook: Threats to privacy. Cambridge, MA: Massachusetts Institute of Technology. [*Unpublished student paper*].
[7] Lile A. and Asilkan Ö. (2011) Analyzing E-Learning Systems Using Data Mining Techniques. *Manuscript ISCIM*