An agent-based model of the development of friendship links within Facebook

SMA Abbas

Centre for Policy Modelling, Manchester Metropolitan University Business School, Aytoun Street, Aytoun Building, M1 3GH, Manchester, UK
ali@cfpm.org

Abstract. This paper investigates how local preferences and social structural constraints might affect the development of the friendship network in Facebook. We do this by analysing a dataset of an American university, Caltech, and by building an agent-based simulation for comparison. Several different, but plausible, processes of friendship network development are proposed in which the structural information of the growing network and the student preferences are taken into account. 'Network formation based on personal preference and social structure' matches the data best, and is thus the preferred hypothesis for the dominant way that friends are added by students in Facebook.

Keywords: Facebook, Social Simulation, SNA, Community Structure

1 Introduction

Since the advent of online Social Networking Systems (SNS), the Internet has changed in terms of its importance to people's everyday lives. It has become part of everyone's life. A huge number of people have a presence over the internet via a "profile", which is a publicly articulated webpage describing a virtual self. Online SNS present themselves as a platform for such profiles. Not only can people present themselves, but can present their social network as well. Since 2004, when Facebook, currently the most popular SNS, came into being, there has been a lot of research on how people form friendships and interact over it. It alone has over 500 Million users to its credit [1].

The magnitude of the data present in the online SNS is enormous, and presents itself as a rich source of social information for analysis. According to a study, most of the online social networks act as a representation of the offline, or real social networks [2]. So it could be assumed as an approximation or a proxy of a real world social network. Not only does an SNS capture the social network, but also the activity between users. But sadly, due to privacy concerns and its commercial value, this data is generally not shared with the research community. So we are left with either a snapshot with limited information, or an activity log without any social network. A huge data set which was touted as the biggest public data of Facebook, is not being

shared due to a technical difficulty [3]. The aim of this paper is to reconstruct the development of the social network with the help of an agent-based methodology, so that a probable history of the social network and an understanding of it could be developed.

A lot of social network based models have been made, but they do not address how such a network might develop within an online environment. This paper attempts to address this concern. First, we simulate a few strategies of how students meet and develop their social network and then we compare the obtained results with the Caltech data set.

In Section 1.2, we discuss the general characteristics of social networks. Then in Section 1.3, we define the data on which our agent-based model is based – its characteristics and network structure. After that, in Section 2, we define our model and the modes of interaction it offers. Simulation results and their comparison with the dataset are presented in Section 3. Related work is summarized in Section 4. At the end, in Section 5, we summarize our findings and present the future outlook of our research by concluding the paper.

1.2 General rules of social network in an SNS

The structure of an SNS can be characterized by its low average distance, moderate clustering coefficient and a power law distribution of number of links [4, 5]. Generally social networks have a moderate clustering coefficient ranging from 0.2 to 0.7, depending on the size and the degree of the network [4] and also a low average distance when compared with a random network with the same density. These results, however, are not proven systematically for all the studied social networks, but nonetheless, could be considered as general guidelines for both online and offline social networks.

1.3 Underlying data

We have used the data of Caltech' students and faculty members. This was provided to us by [6] and has been studied in [7]. The data contains 769 people. It contains eight attributes for each person: ID, student/faculty status, gender, major, second major/minor, dorm/house, year, high school; and also each students friendship links. This dataset only represents intra-institute, in this case, Caltech relationships, which may be the reason why we do not see the average number of friends, as stated by Facebook [1] (130 friends). The data is a snapshot – it represents only links and attributes present at one single point of time. The data is completely anonymized where simple integer values represent each attribute. Although we have included all the eight attributes, but for analysis, we considered only the following four attributes:

- Dormitory
- Major
- Year
- High School

In total, there are 769 people in the dataset. Out of them, 501 have all the values for each attribute. And the total number of links between them is 16656. Missing information in the data is represented by the value 0. In the table below, number of missing values for each attribute is represented.

Table 1. Missing values for each four attributes

Caltech Dataset	Missing Dorm	Missing Major	Missing Year	Missing High School
	172	77	114	134

2 Model Outline

In order to understand the dynamics of social network, we simulate it using an agent-based model. The main aim of the paper is to understand the interplay of social processes and their impact on the network structure as a whole. Thus the key focus is on analyzing how students interact and build their social network over time. Once made, we see which mode of interaction seems to produce the best representation of a social network as judged by a comparison with the underlying dataset. The number of agents is based on the underlying dataset of Caltech University students.

2.1 Rules

In this Section, we discuss how the agents might interact with each other, in terms of making friends in real life and then who adds whom to their list of friends within facebook. We have devised four different plausible strategies of student interactions which are defined below. All agents are initialized with the attributes (major, dorm etc.) of those in the Caltech data set.

1. **Mode 1 - Random Mode**

Each student meets up with a new randomly chosen student after every time step or simulation tick. The target student is selected using a uniform probability distribution. After the selection, the source student determines if the target student satisfies its personal preference. If it does, an undirected link is created among them, which shows that they are friends.

2. **Mode 2 - Friend of a friend Mode**

In this mode, initially all students are asked to make only limited random friends based again on a uniform distribution. This should satisfy both the source and target students' preference. If they do not satisfy, they do not form a link. After this initial phase, personal preferences are not taken into account. Now the new friends are

selected in a "friends-of-friends" manner. In order to best identify who would be the best candidate for it, we select the student with the highest number of friends (who is not already a friend) – which shows how popular she is. This addition goes on, till friends-of-friends are fully explored. After this phase, third degree friends are explored in the same fashion. And when there are no third degree friends left, we apply the same mechanism on the fourth degree friends.

3. Mode 3 - Party Mode

In this mode the personal preferences are also not taken into account. All students arrange a small party which is held on a regular basis. The number of participants in a party is 10. And at each party, 30 new (random) friendships are made. When a party happens, on average 10 friendships are made.

4. Mode 4 - Hybrid Mode

This mode combines the first three modes. At every simulation time step, a simulation mode between 1 and 2 is chosen on a uniform basis. In order not to overwhelm the randomness, Mode 3 is run in every 20^{th} time step. So this mode, essentially, uses all the other three modes.

We do not claim that we present an exhaustive list of modes of interaction. The idea was to come up with some realistic modes which try to capture the micro level preference of agents. This list, by no means, is comprehensive in nature.

Personal preference is not taken into account when there are missing values for all the four attributes. Hence, in this case, we totally neglect the preference of both the source and the target students.

The values for each of the four attributes can be seen in Table 2. All of the four interaction modes use these values. Each of them runs till its network size – the total number of links – matches with the actual dataset of Caltech.

Table 2. Values of the four attributes for all the modes of interactions

Dorm Preference	Major Preference	Year Preference	High School Preference
90	**30**	**20**	**10**

3 Results

In this Section, we compare the simulation results with the actual dataset. First we compare the global or overall results in Section 3.1 and then in Section 3.2, we discuss the attribute level comparison.

3.1 Global results

In this Section, we compare the structure and the community detection mechanism based on the overall network of the actual dataset with the various simulation modes.

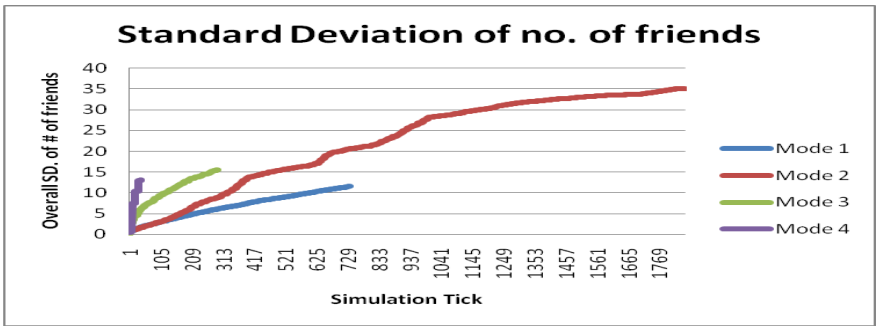

Fig. 1. Standard deviation of the number of friends each agent has in all the four modes

We start off by showing the simulation run of all the four modes. In Figure 1, we have shown how the overall standard deviation in number of friends changes over time using different modes of interactions. As mentioned in Section 2.1, all the modes of interaction run till they reach the same network size as the actual dataset. Some modes take less time than others. Hence we see different end time for each. Modes 1 and 3 have almost linear graph because of their randomness. While Mode 4 being the hybrid mode changes rapidly when Mode 3 is selected and run – producing a hike in number of friends. Mode 2 takes the most simulation ticks and has a high variance. The reason for this is, the network grows depending on the node degree in the neighborhood, which in turn relies on other nodes.

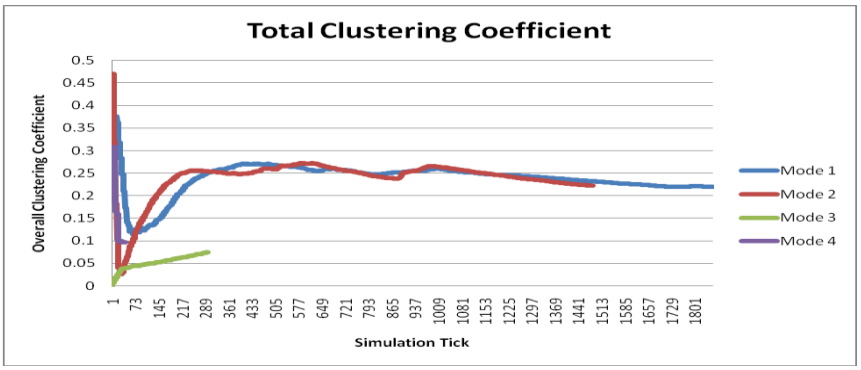

Fig. 2. Total Clustering Coefficient of each mode of interaction

For all the modes of interactions, overall Cluster Coefficient increases in the beginning, as can be seen in Figure 2. The moment it grows out of friends of friends,

it decreases sharply. In Mode 3, however, the links are made total randomly; hence it does not decrease. Mode 4 just combines the effects of all the other three modes – depending on the mode currently being used, it demonstrates the relevent behaviour.

For the selection of the values for each attribute, we relied on statistical measures which were correlations in this case. According to it, the parameter Dorm Preference (DP) plays a significant role in the link developments. Hence we concentrated on it thoroughly to understand the changes of it on the network structure as well as the impact on the attribute based communities. We explored the parameter space for dormitory attribute, starting from 60 to 90 percent preference for the same dorm.

Table 2. Modularity of Mode 1 and Mode 2 with varying Dorm Preference (DP)

Actual	**Modularity - 0.301906**	
Dorm Preference	**Mode 1**	**Mode 2**
90	0.323807	0.322131
80	0.162037	0.16939
70	0.119126	0.121778
60	0.115038	0.11509

As can be seen in Table 2, the closest modularity with the actual dataset is found when the Dorm Preference is set to 90. We have used the method of community modularity as described in [8]. So when the Dorm Preference is set high, the modularity correspondingly also becomes high. Also, in Mode 2, just like Mode 1, the initial random network development which is based on both the source's and the target's preference, acts as a strong characteristic of high modularity network.

Table 3. Fitted centrality degree distribution with varying Dorm Preference (DP)

Actual Dataset	**Normal Distribution - Mean = 0.0282 and Variance = 0.0241**			
Dorm Preference (DP)	**Mode 1 Normal Distribution Parameter Values**		**Mode 2 Normal Distribution Parameter Values**	
	Mean	**Variance**	**Mean**	**Variance**
90	0.028	0.0076	0.028	0.022
80	0.028	0.0055	0.028	0.022
70	0.028	0.0044	0.028	0.025
60	0.028	0.0039	0.028	0.024

In Table 3, we summarize the underlying distribution for the varying Dorm Preference of both Modes 1 and 2. In order to identify the underlying degree distribution, we used the method of Least Square Error (LSE) – the lower the value, the better the fit. And to identify the parameter values for the distribution, we used the method of Maximum Likelihood Estimation (MLE). Although the underlying distribution of the actual dataset and Mode 2 with DP being 90 were Beta Distributions when Least Squared Method (LSM) was applied to them, but with a

very minor difference, Normal Distribution was also a good fit. And since most of the simulation results of both the modes reveal that they are Normal in nature, we considered Normal Distribution. There is a major difference between the two modes. In the case of Mode 1, the variance decreases as the DP is decreased, while Mode 2 shows almost similar behavior in all the variable DP values. It can be said that there is a very low impact on network structure of initial friendships in Mode 2 which are based on personal preferences. We are focused on both community and network structure, hence we select DP to be 90, as it is a better candidate for network modularity. From now on, by default the value of DP is 90 for all the modes of interactions.

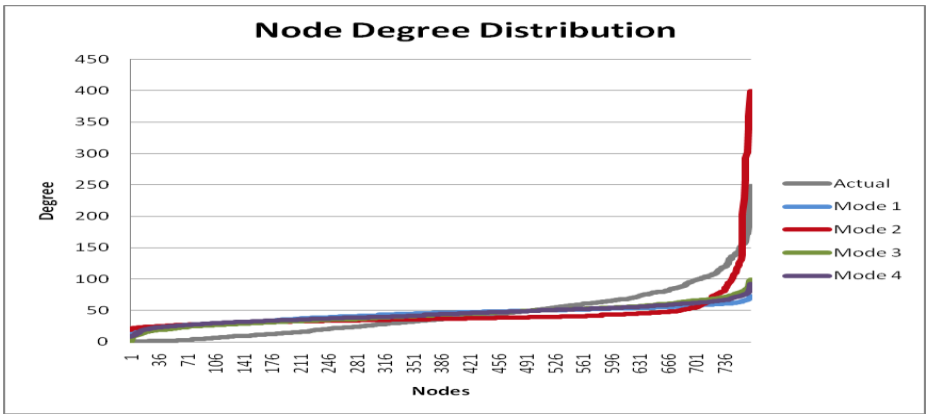

Fig. 3. Degree distribution of all the four simulation modes and the actual dataset.

We have summarized in Figure 3, the degree distribution of the actual and the four interaction modes. This only shows the final node degrees after the simulation has been finished. The actual and the Mode 2 degree distributions show a power law effect which suggests that most of the nodes have few links while only a few nodes have a lot of links. The other three modes, Mode 1, 3 and 4 seem *normal* in nature. Their links are more or less uniformly distributed.

We have concentrated on a few and important factors of Social Network Analysis (SNA) in order to compare the actual with the simulated network. The factors with their respective values can be seen in Table 4:

Table 4. Important SNA attributes of the actual and the modes of interactions between agents

Model Type	Avg. Distance	Connectedness	Cluster Coefficient	SD. of number of friends	Community Modularity
Actual	**2.4747**	**0.98**	**0.23**	**37.03**	**0.301906**
Mode 1	2.4929	1	0.219	11.52	0.323807
Mode 2	**2.6187**	**1**	**0.222**	**35.05**	**0.322131**
Mode 3	2.3909	1	0.074	15.55	0.117925
Mode 4	2.4906	1	0.09	13.71	0.126717

In Table 4 we can clearly identify that Mode 2 remains the best candidate when it is compared with the actual dataset. Although the actual dataset is not a fully connected network, but the average distance, the standard deviation of number of friends, total cluster coefficient and even the overall modularity is quite similar to the actual social network. The underlying distribution of both the actual and Mode 2 can be identified by such a huge standard deviation; which in turn reflects our earlier finding that both of these are in fact power law distribution.

3.2 Attribute Level Results

In this Section, we compare the results of our simulation runs of all the four modes for each of the attributes with the actual dataset. We measured the results in terms of *Silo Index*. This is an Index which identifies the degree of inter-links between nodes with a particular attribute in a (social) network. If an attribute *X* has all the links to itself, and not to any other attribute, that means a very strong community exists, which is totally disconnected from the rest of the network. In short, this index helps us identify how cohesive inter-attribute links are. It ranges from *-1* to *1*, representing the extreme cases (no in-group links to only in-group links respectively). We have presented our results in Figures 4-7 below.

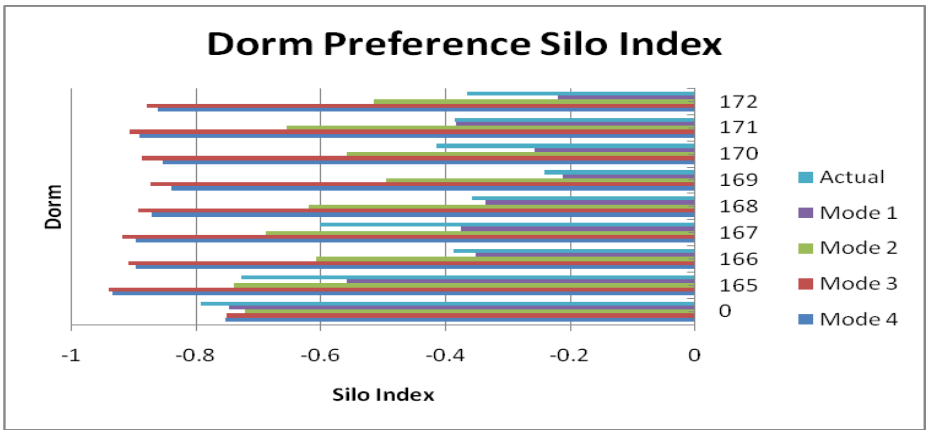

Fig. 4. Silo Index for Dorm Preference for all the four modes and the actual network

In Figure 4, we calculate Silo Index of the Dorm Preference (DP) attribute. This method was run on all the four modes and the actual dataset. If we see the difference of each mode to the actual dataset, Mode 1 has the least difference. Then Mode 2, 3 and 4 come according to their differences with the dataset. There is one interesting thing to be noticed here. Since randomness in Mode 4 is introduced by Mode 3, it resembles a lot with it.

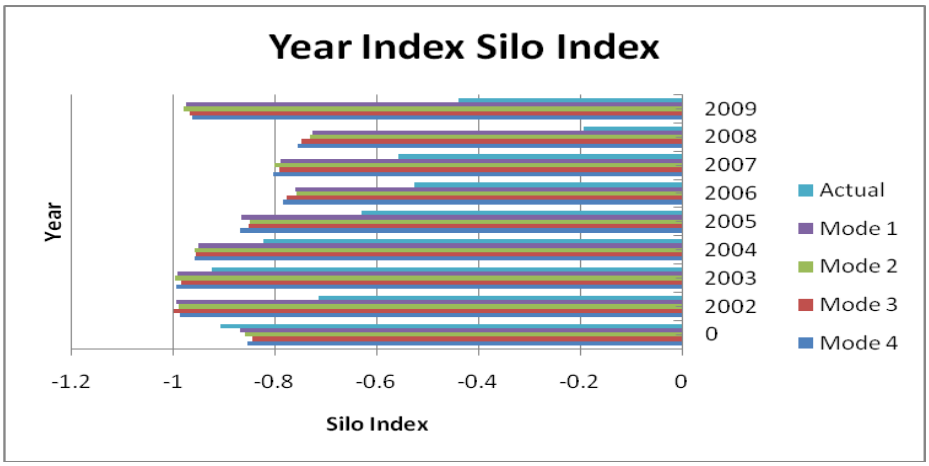

Fig. 5. Silo Index for Year Preference for all the four modes and the actual network

In the case of year, as can be seen in Figure 5, almost all the modes behave similarly. The reason being the insignificant correlation of the Year Preference (YP) in the actual dataset that random and preferential attachments do not vary that much.

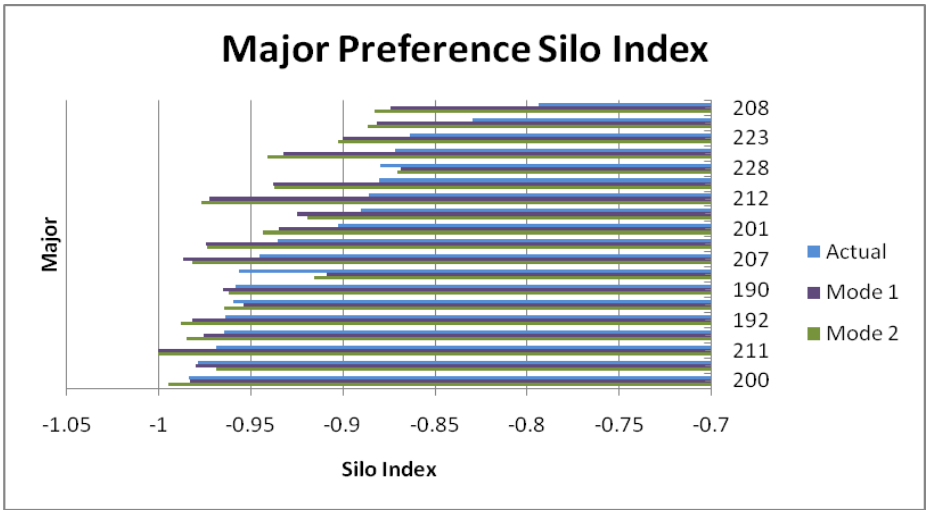

Fig. 6. Silo Index for Major Preference (MP) for all the four modes and the actual network (showing IDs of the most popular majors only)

Since the results so far have favoured the first two modes, Modes 1 and 2, we now only focus on them for our next set of results. In Figure 6, the Silo Index of the Major

Preference (MP) is shown. Both Modes 1 and 2 have a minor difference with the Silo Index of the actual dataset and are quite similar to each other.

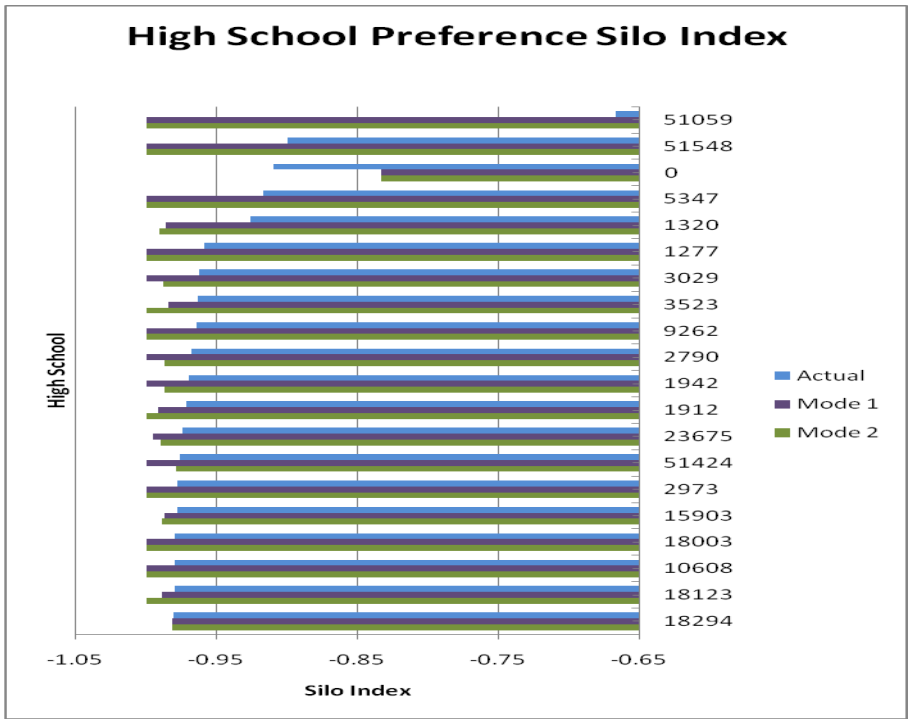

Fig. 7. Silo Index for High School Preference (HSP) for all the four modes and the actual network (showing IDs of the most popular high schools only)

The High School Preference (HSP) also has very insignificant correlation – hence very low Silo Index. In Figure 7, we can see that both Modes 1 and 2 have similar behavior and can be considered good representation of the actual dataset.

After comparing all the four attributes, Mode 1 takes the lead in the DP, but for the other three attributes, Modes 1 and 2 both present themselves as good candidates.

4 Related Work

A plethora of research in SNS has been done over the last five year. The major focus of such works has been the identification of the static nature of SNS [9].

Adalbert studied Facebook from an economist's point of view [10]. The data which he collected and then studied showed that race plays the most significant role in student friendship development – especially in the case of minority. In an another study carried out on students' network [11], race and local proximity, such as dorm were determined to play the most important role, followed by common interests such

as major and similar social standing, which in turn were followed by common characteristics such as same year. In our data, however, we could not verify the race factor, as this information is not present in the dataset that we have used. In case of SNS growth, unlike our model, there are some studies that identify the different classes of users [12]. And also, based on the activity of users, a couple of studies show their social network development [13]. Based on only the structure of an SNS, a couple of exploration techniques have also been devised to predict what new links users are going to make [14, 15], but they usually do not take into account the rich information of attributes of users [16].

5 Conclusion and Future Outlook

Our agent-based model tries to develop an understanding of students' choice in link developments. Both endogenous and exogenous factors have been taken into consideration.

This is a preliminary work in which we tried to understand how local preferences and the structural factors help develop a social network. We have devised a few interaction strategies of interaction between students. We compared our simulation outcomes to the students' network of Caltech University. We relied on both community detection method and major SNA factors for comparison. The strategies of interaction varied from preferential attachment – based on the attribute values, to complete random interactions. In the hybrid mode i.e. Mode 4, the randomness of mode 3 has a major influence on it so we did not see much difference among the two modes – be it general or attribute level comparison.

After analyzing the results and comparing them with the actual dataset, we determined that Mode 2, which initially takes local preferences into account but then works on a Friend-of-a-friend basis, does the best. It captures the basic essence of the underlying network. From network level measures to the attribute level comparison, it presents itself as a good candidate for the understanding of students' interactions and social network development. The initial setting of highly similar friends leads to a cohesive community structure and also the friends-of-a-friend process with has a power law outlook. Modes 3 and 4 which are dominated by the random meeting of friends at events did not explain the data well.

We do not claim that we presented an exhaustive list of social processes, but analyzed a few possible kinds of interaction. Focusing on personal preference and on social structure, presents itself as a promising mode of interaction. While only pre-simulation statistics, such as Correlation, do not necessarily present the best parameter values. For the initial friendship links, the parameter space has to be explored to find the best match.

In future, we would like to make a more general model, which captures both local and global aspects of a social network. This model will be based on several datasets and on the findings of this model. Also, with the aid of the earlier studies on social network - specifically online social network, we will try to design and understand the processes involved. We will focus both on internal and environmental aspects.

Acknowledgments

We would like to thank Bruce Edmonds of Centre for Policy Modelling, for his feedback and useful suggestions.

References

1. http://www.facebook.com/press/info.php?statistics
2. Ellison, N. B., Steinfield, C. and Lampe, C. (2007), The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. Journal of Computer-Mediated Communication, 12: 1143–1168. doi: 10.1111/j.1083-6101.2007.00367.x.
3. Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, *30*(4), 330-342. doi: 10.1016/j.socnet.2008.07.002.
4. Dekker, AH, Abstract, E. (2004). Realistic Social Networks for Simulation using Network Rewiring. *October*, (i), 677-683.
5. Newman, M. E. J. (2000). Power-law distributions in empirical data. *Physics*.
6. http://people.maths.ox.ac.uk/~porterm/data/facebook5.zip
7. Traud, A. L., Kelsic, E. D., Mucha, P. J., Porter, M. A., Interdisciplinary, F. O. R., Mathematics, A., et al. (n.d.). Community structure in online collegiate social networks. *North*, 1-15.
8. Finding community structure in very large networks, Aaron Clauset, M. E. J. Newman, and Cristopher Moore, Phys. Rev. E 70, 066111 (2004).
9. Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC '07*, 29. New York, New York, USA: ACM Press. doi: 10.1145/1298306.1298311.
10. Mayer, A. (2009). Online social networks in economics. *Decision Support Systems*, *47*(3), 169-184. Elsevier B.V. doi: 10.1016/j.dss.2009.02.009.
11. B. Sacerdote, Mararmos, How do friendships form? The Quarterly Journal of Economics 121 (1) (2006).
12. Kumar, R., Novak, J., Tomkins, A., 2006. Structure and evolution of online social networks, in: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA. pp. 611–617.
13. Golder, S.A., Wilkinson, D., Huberman, B.A., 2007. Rhythms of social interaction: Messaging within a massive online network, in: Steinfield, C., Pentland, B., Ackerman, M., Contractor, N. (Eds.), Proceedings of Third International Conference on Communities and Technologies. Springer, London, U.K., pp. 41–66.
14. L. Backstrom and J. Leskovec. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In the proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM), pages 635-644, 2011.
15. A. Agarwal and S. Chakrabarti. Learning random walks to rank nodes in graphs. In the proceedings of the 24th International Conference on Machine Learning (ICML), pages 9-16, 2007.
16. Gao, B., & Wang, T. (n.d.). Semi-Supervised Ranking on Very Large Graph with Rich Metadata. *Machine Learning*, (49).