

# 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](mailto: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 way that students add “friends” on Facebook.

**Keywords:** Facebook, Social Simulation, SNA, Community Structure

## 1 Introduction

Since the advent of online Social Networking Systems (SNS), the Internet has become part of everyone’s everyday 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, e.g. [3][7][16]. Facebook alone has over 750 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 of longitudinal nature of Facebook has been collected, but with a limited access [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 possible history of the social network and an understanding of it could be developed.

A lot of social network based models have been made. From a general but realistic social network (e.g. see [4][5]) to a data-driven students' social network[6], 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 some possible strategies of how students meet and develop their social network. Then we compare the obtained results with the underlying dataset we have used and in this way are able to make some inferences as to the probable strategies that the students used.

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 characteristics of SNS “friendship” networks**

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 [7, 8]. Generally such social networks have a moderate clustering coefficient ranging from 0.2 to 0.7, depending on the size and the degree of the network [7] 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 social networks, but nonetheless, could be considered as general guidelines for them.

## **1.3 The reference data**

We have used the data of students and faculty members of Caltech who use Facebook. This was provided to us by Mason A. Porter of Oxford University, and has been studied by him and others in [9]. The dataset includes both the attributes and social structure for 769 people. For each person, it contains eight attributes, which are: 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 relationships, which may be the reason why we do not see the average number of friends, as stated by Facebook [1] (130 friends). It 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: their dormitory; their year (1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, etc.); their “major” (their main subject of study); and the high school they came from.

Out of the 769 people in the dataset, 501 have all the values for each attribute. And the total number of links between them is 16656. Table 1 summarizes the extent of missing data (indicated by a “0”).

**Table 1.** Summary of missing values (for the four used attributes)

<b>Caltech Dataset</b>	<b>Missing Dorm</b>	<b>Missing Major</b>	<b>Missing Year</b>	<b>Missing High School</b>
	172	77	114	134

## 2 Model Outline

In order to understand the dynamics of this social network, we simulate it using an agent-based simulation. 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. We can then see which mode of interaction seems to produce the best representation of a social network as judged by a comparison with the reference dataset. In this section, the term *agent* will be used to refer to student.

### 2.2 Simulation Setup, Execution and Termination

The number of agents in all simulation runs is 769, based on the underlying dataset of Caltech University students. Each individual in the dataset was provided the attributes for one agent in the simulation. All agents are created at the start. While initializing a simulation run, the order of the agents is not taken in any way. They are randomly chosen. Interaction strategy for all the agents is set once in the beginning. It does not change. Each simulation runs until the number of links made is the same as in the reference dataset – 16656. No link is dropped or modified once it is created.

### 2.1 Rules for Adding Friends

In this Section, we discuss how the agents might interact with each other, in terms of making friends in real life. It is assumed that, by and large, these real life social links will then be duplicated within Facebook. We do not claim that we present an exhaustive list of possible strategies; rather the idea is to explore *some* plausible ways that depend on the micro-level preference of agents and then evaluate them.

Each agent is initialized with the four attributes (major, dorm etc.) of a corresponding individual recorded in the Caltech data set. The values for each of the four attributes can be seen in Table 2.

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

<b>Dorm Preference</b>	<b>Major Preference</b>	<b>Year Preference</b>	<b>High School Preference</b>
<b>90</b>	<b>30</b>	<b>20</b>	<b>10</b>

All agents have a preference for each of the four attributes we have selected which is known as “*Personal Preference*”. The idea has been inspired from homophily – the love of the similar [10]. It is a probabilistic match of attributes between the source and the target agents. We have shown the illustration in Table 3 for the dormitory attribute. A *chance* out of 100 is randomly selected in a uniform fashion. If it is under the predefined preference value (90 in case of dormitory preference) and the attribute values of both the source and target agents are known and match with each other, then the dormitory preference is satisfied; and we set the dormitory flag to true. Also, if the *chance* is greater than the preference value, it is satisfied as well. We repeat the same process for the remaining attributes. If all the four attributes’ conditions are satisfied, we make a friendship link between the source and the target agents.

**Table 3.** Algorithm to calculate “*Personal Preference*”

```

1. Agent Source = getSourceAgent(), Agent Target = getTargetAgent()
2. Integer DP = getDPValue() // Get dormitory preference value which is fixed as 90
3. Boolean sameDorm = False, Integer chance = get_random_integer(100)
4. IF (chance < DP){ // 0<chance<=DP
5.     IF (Source.getDorm() == Target.getDorm()) AND
6.         (Source.getDorm() != 0 And Target.getDorm() != 0){ sameDorm = True }
7. }ELSE{ sameDorm = True }
8. ...
9. //repeat the same evaluation for the rest of the attributes (Major, Year etc)
10. IF (sameDorm AND sameMajor AND sameYear AND sameHighSchool)
    // If all conditions satisfy
11.     form_a_link(Source, Target) //create a friendship link between the two

```

We have devised four different plausible strategies (here called “modes”) for agent interaction – each involves matching students using their attributes, but in different ways. 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 agent. All of the four interaction modes use the attribute values defined in Table 2.

### 1. Mode 1 - Random Mode

Each source agent selects a randomly chosen target agent after every time step or simulation tick. The target agent is selected using a uniform probability distribution. After the selection, the source agent determines if the target agent 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, there are two phases. In the first phase, all agents are asked to make only limited random friends selected in a uniform distribution. This should satisfy both the source and target agents' preference. If they do not satisfy, they do not form a link. After this initial phase, personal preferences are not taken into account. From then on, in the second phase, new friends are selected in a "friends-of-friends" manner. During this phase, at each iteration, based on a snapshot of one's social network, each agent selects a friend-of-a-friend with the highest number of friends – which shows how popular they are. If no link exists between them, we create a link. And if they are already connected (a cycle exists), third degree friends are explored in the same fashion. And when there are no link exist between the highest third degree agent and the self, the same mechanism applies to 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. The selection of the party participants is totally independent and unbiased towards any attribute. At each party, a maximum of 30 new (random) friendships are made. Due to the random selection of party participants, there is a chance of selecting nodes which are already connected to each other. In that case, no new link is established.

### **4. Mode 4 - Hybrid Mode**

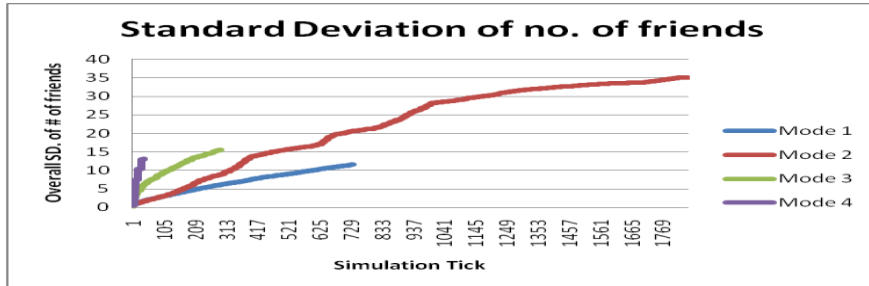
This mode is a combination of the above three modes. At every simulation time step, a simulation mode between 1 and 2 is chosen on a uniform basis. In order not to overwhelm the randomness, Mode 3 is run in every 20<sup>th</sup> time step.

## **3 Results**

In this Section, we compare the simulation results with the reference 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.

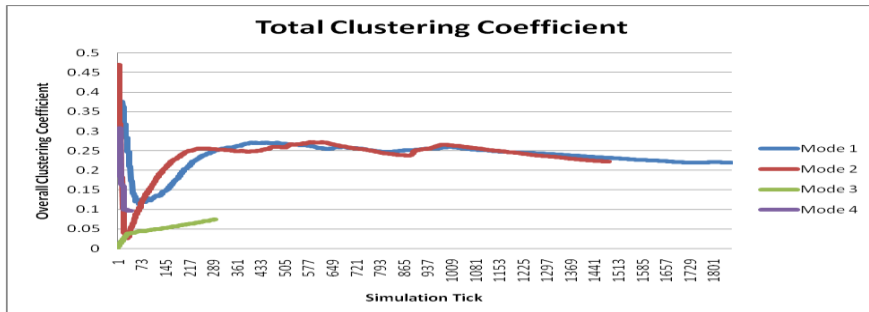
### **3.1 Global results**

In this Section, we compare the structure and the community detection mechanism based on the overall network of the reference 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, the simulation using any of the modes of interaction terminate when they reach the same network size as the reference dataset. Some modes take less time to do this 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 relevant 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 4.** Modularity of Mode 1 and Mode 2 with varying Dorm Preference (DP)

Reference	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

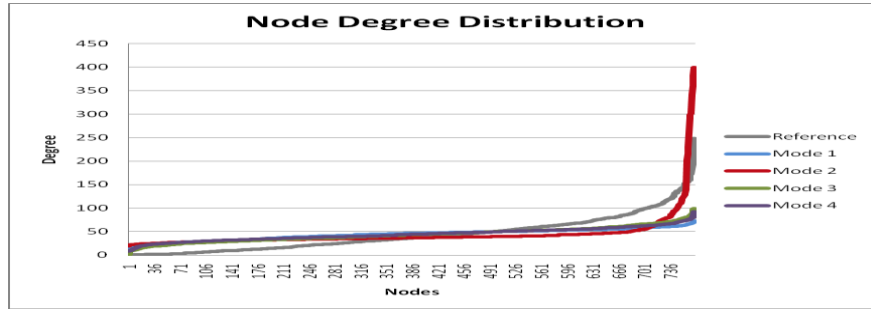
As can be seen in Table 4, the closest modularity with the reference dataset is found when the Dorm Preference is set to 90. We have used the method of community modularity as described in [11]. So when the Dorm Preference is set high, the modularity correspondingly also becomes high. Also, in Mode 2, 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 5.** Fitted centrality degree distribution with varying Dorm Preference (DP)

Reference 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 5, 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 reference 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, a DP value of 90 is used in all the following results.



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

We have summarized in Figure 3, the degree distribution of the reference and the four interaction modes. This only shows the final node degrees after the simulation has been finished. The reference 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 reference with the simulated network. The factors with their respective values can be seen in Table 6:

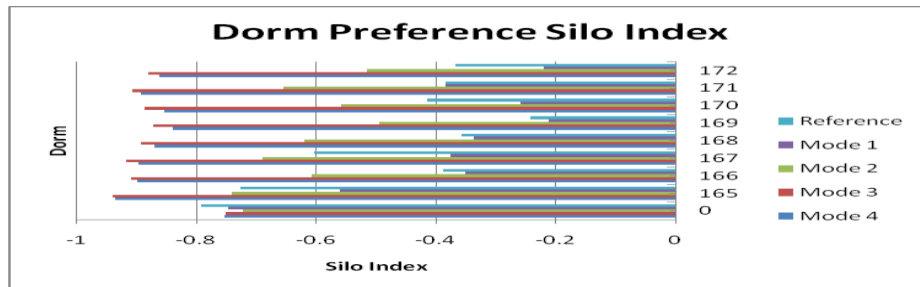
**Table 6.** Important SNA attributes of the reference and the modes of interactions between agents

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

In Table 6 we can clearly identify that Mode 2 remains the best candidate when it is compared with the reference dataset. Although the reference 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 reference social network. The underlying distribution of both the reference 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.

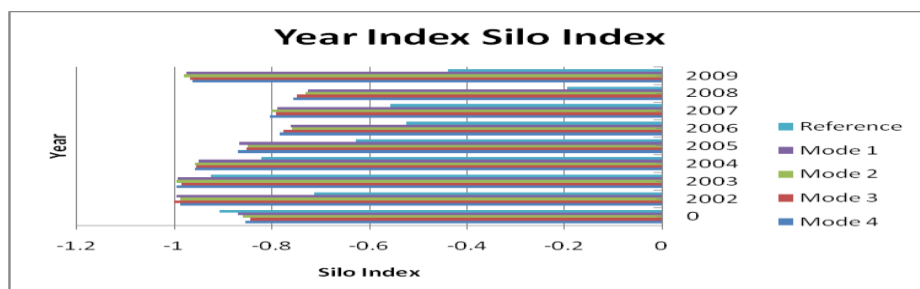
### 3.2 Attribute Level Results

In this Section, we compare the results of our simulation runs of all the four modes for each of the attributes with the reference dataset. We measured the results in terms of the *Silo Index*. This is an Index which identifies the degree of inter-links between nodes with a particular attribute value in a (social) network. If a set of nodes having a value  $Y$  for an attribute  $X$ , has all the links to itself, and not to any other values of attribute  $X$ , 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 reference network

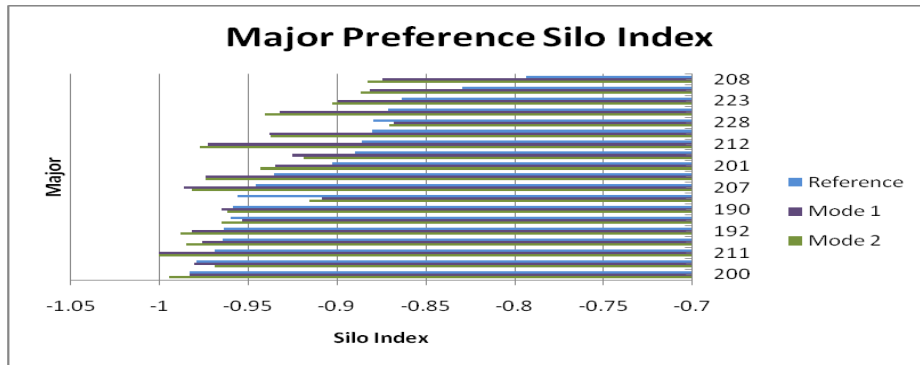
In Figure 4, we calculate Silo Index of the Dorm Preference (DP) attribute. This method was run on all the four modes and the reference dataset. If we see the difference of each mode to the reference dataset, Mode 1 has the least difference. Then Mode 2, 3 and 4 come according to their differences with 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 reference network

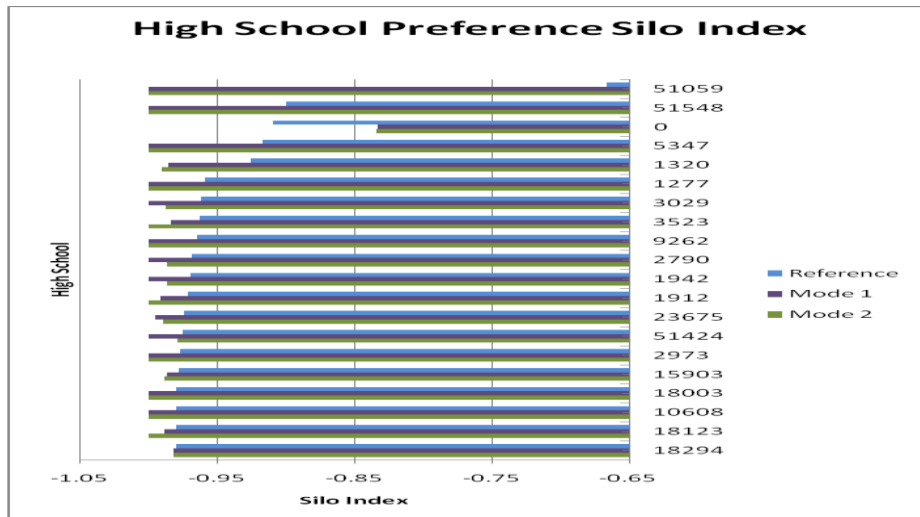
In the case of year, as can be seen in Figure 5, almost all the modes behave similarly. The reason being the insignificant correlation of the Year Preference (YP)

in the reference 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 reference 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 reference dataset and are quite similar to each other.



**Fig. 7.** Silo Index for High School Preference (HSP) for all the four modes and the reference 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 reference 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. It is impossible to cover all of it; hence some of the relevant work is being mentioned here. The major focus of such work has been the identification of the static nature of SNS (e.g. [13]).

To understand the behavior of students' real social network development, a function of contact frequency and shared interests has been used in [6] to make a model. Jackson et al. in [12] developed a model in which a neighbourhood search is done to develop a social network; this can result in many of the characteristics of observed networks.

Adalbert studied Facebook from an economist's point of view [14]. 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 his previous study[15], out of students of Texas A&M, he found that majority of meeting new friends (26%), were driven by members of the same school organizations. In an another study carried out on students' network [16], 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 [17]. And also, based on the activity of users, a couple of studies show their social network development [18]. 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 [19, 20], but they usually do not take into account the rich information of attributes of users [21].

## 5 Conclusion and Future Outlook

An agent-based simulation has been described that attempts to explain how students make SNS links, taking into account both endogenous and exogenous factors.

This is a preliminary work in which we tried to understand how local preferences and the structural factors might help develop a social network. We have devised and explored a limited number of strategies for student interaction. We compared our simulation outcomes to data gathered from students' Facebook 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. We did try to control Mode 3 selection in this mode, but the randomness of Mode 1 also did not quite help in the social network development. Although the attribute level communities produce comparable results with the reference network of the dataset, but the total random selection of target nodes in Mode 1 resulted in low overall cluster coefficient and low standard deviation in number of friends.

After analyzing the results and comparing them with the reference 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 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 possible social processes, but rather analyzed a few plausible variations. Focusing on personal preference and on social structure, presents itself as a promising mode of interaction. While only pre-simulation statistics based on the underlying data, 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. We are in the process of collecting longitudinal data to better inform this. 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 the Manchester Metropolitan University Business School for the studentship under which this research was done, to Mason Porter of Oxford University for providing us the Caltech dataset and 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.
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.
4. Hamill, L Gilbert, N., (2006). A Simple but More Realistic Agent-based Model of a Social Network. *Center for Research in Social Simulation*.

5. Hamill, L. (2010). Communications , Travel and Social Networks since 1840: A Study Using Agent-based Models. *Social Networks*.
6. Singer, H. M., Singer, I., & Herrmann, H. J. (2009). Agent-based model for friendship in social networks. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 80(2 Pt 2), 026113. <http://www.ncbi.nlm.nih.gov/pubmed/19792206>.
7. Dekker, AH, Abstract, E. (2004). Realistic Social Networks for Simulation using Network Rewiring. *October*, (i), 677-683.
8. Newman, M. E. J. (2000). Power-law distributions in empirical data. *Physics*.
9. 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.
10. McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1), 415-444.
11. Aaron C., M. E. J. Newman, and Cristopher M.(2004). Finding community structure in very large networks, *Phys. Rev. E* 70, 066111 (2004).
12. Jackson, M. O., & Rogers, B. W. (2007). Meeting Strangers and Friends of Friends: How Random Are Social Networks? *American Economic Review*, 97(3), 890-915.
13. 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.
14. Mayer, A. (2009). Online social networks in economics. *Decision Support Systems*, 47(3), 169-184. Elsevier B.V.
15. Mayer, A., Puller, S. L. (2008). The old boy (and girl) network: social network formation on university campuses. *Journal of Public Economics*, 92, 329-347.
16. Sacerdote, B., Marmaros, D. (2006). How do friendships form? *The Quarterly Journal of Economics* 121 (1).
17. Kumar, R., Novak, J., Tomkins, A., (2006). Structure and evolution of online social networks, in: *KDD '06: Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, ACM, New York, NY, USA. pp. 611–617.
18. Golder, S.A., Wilkinson, D., Huberman, B.A., (2007). Rhythms of social interaction: Messaging within a massive online network, in: *Steinfeld, C., Pentland, B., Ackerman, M., Contractor, N. (Eds.), Proc. of 3rd Int. Conf. on Communities and Technologies*. Springer, London, U.K., pp. 41–66.
19. Backstrom, L. and Leskovec, J. (2011). Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *Proc. of the 4th ACM Int. Conf. on Web Search and Data Mining (WSDM)*, pages 635-644.
20. Agarwal, A., Chakrabarti, S. (2007). Learning random walks to rank nodes in graphs. *Proc. of the 24th Int. Conf. on Machine Learning (ICML)*, pages 9-16.
21. Gao, B., Wang, T. (n.d.). (2011). Semi-Supervised Ranking on Very Large Graph with Rich Metadata. *Machine Learning*, (49). *Proc. of the 11<sup>th</sup> ACM SIGKDD Int. Conf. on Knowledge discovery in data mining*.