Network Analysis and Statistics in Investigating Community Structure of High School Texting Network

Oksana Pichugina

Brock University, Department of Mathematics and Statistics

Contact Information:
Department of Mathematics and Statistics
Brock University
500 Glenridge Ave, St Catharines
ON L2S 3A1, Canada
Email: op13ci@brocku.ca

Abstract

The information about texting contacts and their intensity as well as grades, gender, interests and location was collected for about 400 high school students. By means of Statistics and Network Analysis a High School Texting Network was created and analysed in such directions: detecting and naming communities (CD), small world features, the missing data restoring and so on. Several CD algorithms were applied for investigating this network and compared.

Introduction

Network Analysis (NA) is a powerful tool for study complex real-world systems numbering sometimes millions of elements. Graph Theory and Statistics are mathematical instruments for NA. Powerful software (Gephi, IGraph, SNAP) provides computer support of these this analysis. This project was performed in framework of Brock University Mentorship in Science Program and in collaboration with Stephanie Noel and under supervision of professor Babak Farzad.

Main Objectives

For the network of high school students residing in Thorold (Ontario) and the neighbourhood:

1. to collect primarily information;
2. to construct the network;
3. to investigate this network by means of Graph Theory and Mathematical Statistics, in particular:
   (a) to study small-world phenomena;
   (b) to figure out the community structure by several community detection algorithms (CD algorithms, CDA);
   (c) to validate the results by comparison with actual communities;
   (d) to name these communities using the priori and the posteriori information;
   (e) to investigate various CDA effectiveness depending on the type of weighted function.

Data Set

High School Texting Network (HSTN, 521 nodes, 2725 links, weighted) was derived from collected 398 questionnaires (Fig.1)

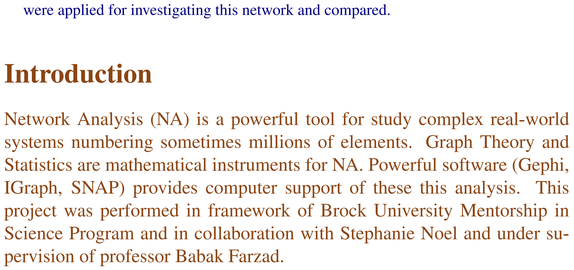

Figure 1: Questionnaire

Small-World Networks (SWN, Watts-Strogatz model)

Definition. A SWN is a network G where the typical distance $d(u, v)$, between two randomly chosen nodes $u, v \in G$, grows proportionally to the logarithm of the number of nodes in the network, that is

$$d(u, v) \propto log\ n, n = |G| \tag{1}$$

Main features of SWN:

1. a high clustering coefficient (CC);
2. a small average shortest path length (ASPL).

Some properties of SWN:

1. cliques, and near-cliques (consequence of a high CC);
2. most pairs of nodes will be connected by at least one short path (consequence of small ASPL).

Several other (typical) properties:

1. an over-abundance of hubs (high degree nodes);
2. a power law degree-distribution (scale-free networks);
3. clear community structure and as a consequence high modularity.

Clustering Coefficient (CC)

Definition. The local CC (LCC) C_i for a vertex v_i is the proportion of links between the vertices within its neighbourhood (n_i) divided by the number of links that could possibly exist between them

$$C_i = \frac{2n_i}{k_i(k_i - 1)}, k_i = |N(k_i)| \tag{2}$$

Definition. The average CC (ACC) $\overline{C}$ is the mean of the LCC C_i of all the vertices:

$$\overline{C} = \frac{1}{|G|}\sum_i C_i \tag{3}$$

Definition. The shortest path length (SPL) $l(u, v)$ between vertices $u, v \in G$ is minimal sum of the weights of its constituent edges.

Definition. The average SPL (ASPL) $\overline{l}$ is the mean of the SPL, averaged over all pairs of nodes.

Definition. Modularity Q (function) is the fraction of the edges that fall within the given groups minus expected value of such a fraction if edges were randomly distributed.

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \tag{4}$$

HSTN Analysis in Gephi

Characteristics	The Texting Network	Random Graph	Prorortion (II/I)
Number of nodes	521	521	100%
Number of edges	2725	2613	96%
Average degree	5,23	5,015	96%
Weighted Average degree	9,234	5,015	
Diameter	8	5	63%
Radius	0	4	
Average Path length	3,841	2,953	77%
The shortest paths range	0-8	4-5	
Number of shortest paths	256544	270920	106%
Modularity	0,632	0,262	41%
Number of Communities	22	12	55%
Number of Connected Components	14	1	7%
Average Clustering Coefficient	0,257	0,019	7%
Total triangles	1339	178	13%

Table 1: Metrics Summary

Community	# of members	Cumulative Portion
4	104	20%
2	103	40%
10	75	54%
20	71	68%
1	60	79%
	41	87%
8	32	
3	14	
7	7	
6	2	
Another 12 communities	1	
Overall	521	87%

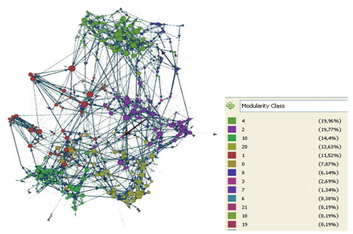

Figure 2: Community Detection Summary

Comparison of detecting communities with actual ones. Calling first 6 communities

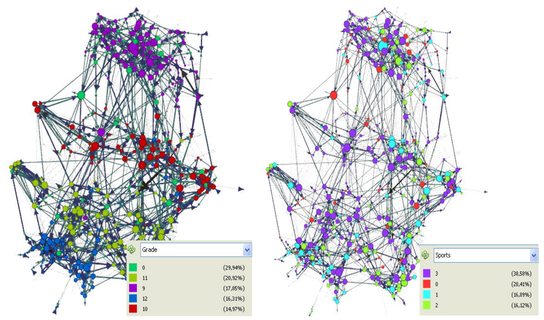

Figure 3: Observing Actual Communities "Sports", "Grade"

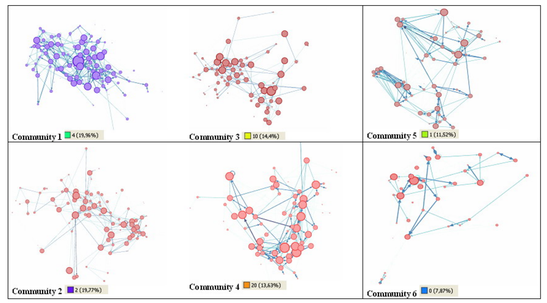

Figure 4: The Communities 1-6 (87 percentages of nodes)

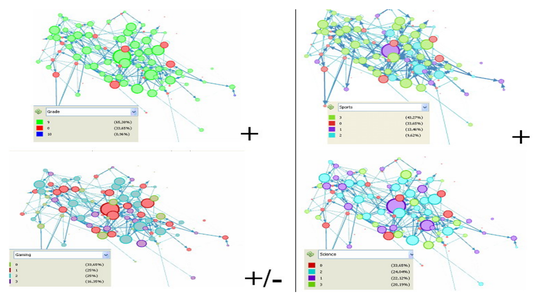

Figure 5: The Community 1 Grade 9, Sport +,Gaming +/-

Community	Grade				Gender		Sports			Science			Gaming		
	9	10	11	12	M	F	1	2	3	1	2	3	1	2	3
1															
2															
3															
4															
5															
6															
Sample Proportion	25%	21%	30%	23%	55%	45%	24%	23%	54%	33%	37%	30%	45%	27%	27%

Table 2: Community Detection Result by Louvain Modularity Method

HSTN Analisis in IGraph

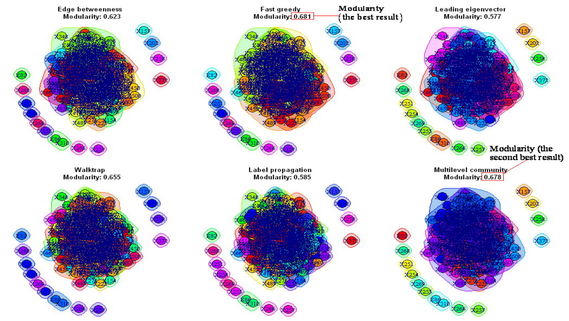

Figure 6: IGraph CD. The highest observed modularity (par=2)

CD algorithm	#	unweighted	weighted						weighted	
		nw	1	1,256	1,538	2	2,462	2,734	3	min	max
Edge betweenness	1	0,606	0,618	0,627	0,628	0,63	0,637	0,631	0,64	0,618	0,638
Fast greedy	2	0,607	0,652	0,648	0,645	0,642	0,619	0,626	0,64	0,619	0,652
Leading eigenvector	3	0,554	0,589	0,598	0,597	0,57	0,594	0,604	0,52	0,519	0,604
Walktrap	4	0,597	0,625	0,621	0,621	0,612	0,618	0,618	0,62	0,612	0,625
Label propagation	5	0,609	0,585	0,584	0,579	0,601	0,59	0,589	0,61	0,579	0,608
Multilevel community	6	0,622	0,65	0,643	0,643	0,645	0,645	0,639	0,65	0,639	0,65
Average		**0,599**	**0,620**	**0,620**	**0,619**	**0,617**	**0,617**	**0,618**	**0,611**	**0,611**	**0,620**

Table 3: The worst and the best observed modularity for different CDA

CD algorithm	#	unweighted	weighted						weighted
		nw	1	1,256	1,538	2	2,462	2,734	3	Average
Edge betweenness	1	0,606	0,618	0,627	0,628	0,63	0,637	0,631	0,64	0,630
Fast greedy	2	0,607	0,652	0,648	0,645	0,642	0,619	0,626	0,64	0,639
Leading eigenvector	3	0,554	0,589	0,598	0,597	0,57	0,594	0,604	0,52	0,582
Walktrap	4	0,597	0,625	0,621	0,621	0,612	0,618	0,618	0,62	0,619
Label propagation	5	0,609	0,585	0,584	0,579	0,601	0,59	0,589	0,61	0,591
Multilevel community	6	0,622	0,65	0,643	0,643	0,645	0,645	0,639	0,65	0,644
min		**0,554**	**0,585**	**0,584**	**0,579**	**0,570**	**0,590**	**0,589**	**0,519**	**0,582**
max		**0,622**	**0,652**	**0,648**	**0,645**	**0,645**	**0,645**	**0,639**	**0,646**	**0,644**

Table 4: The worst and the best observed modularity for different weights

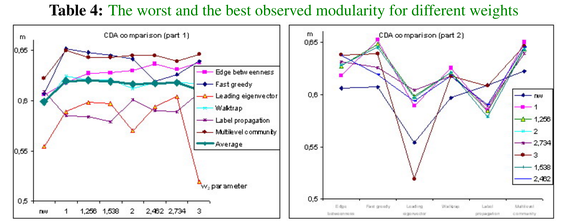

Figure 7: Dependency of modularity on CDA and weights

Conclusions

The High School Texting Network for Thorold area was derived and investigated by means of Graph Theory and Statistics. As it was expected for this network all Small-World features are valid. Accuracy of Gephi and Igraph community detection (CD) was explicitly verified by comparison with actual communities. Feasibility of using weighted graphs for CD was confirmed. In addition, dependency of modularity from weights choice was investigated.