Analyzing the Facebook Friendship Graph

Salvatore Catanese[1], Pasquale De Meo[2], Emilio Ferrara[3], Giacomo Fiumara[1]

[1] Dept. of Physics, Informatics Section. University of Messina.
`salvocatanese@gmail.com; giacomo.fiumara@unime.it`
[2] Dept. of Computer Sciences, Vrije Universiteit Amsterdam.
`pasquale.de.meo@few.vu.nl`
[3] Dept. of Mathematics. University of Messina.
`emilio.ferrara@unime.it`

Abstract. Online Social Networks (OSN) during last years acquired a huge and increasing popularity as one of the most important emerging Web phenomena, deeply modifying the behavior of users and contributing to build a solid substrate of connections and relationships among people using the Web. In this preliminary work paper, our purpose is to analyze Facebook, considering a significant sample of data reflecting relationships among subscribed users. Our goal is to extract, from this platform, relevant information about the distribution of these relations and exploit tools and algorithms provided by the Social Network Analysis (SNA) to discover and, possibly, understand underlying similarities between the developing of OSN and real-life social networks.

Keywords: social networks, analysis, visualization, graphs, data mining

1 Introduction

The problem of analyzing social networks was already introduced during late sixties by Milgram [18] and Travers [23] in Psychology and Sociology. Starting from this point for twenty and more years, several kind of real-life social experiments have been conducted and studied by sociologists, trying to understand motivations, dynamics and rules of real-life social networks.

During last years the Web phenomenon of OSN started spreading and computational aspects have also been considered [3, 15, 22]. Several Social Networking Services were developed, most of them gathered millions of users in an incredible short amount of time. The OSN we are considering in this work is Facebook [4], which collected more than 500 millions of world-wide users as of July 2010.

The unpredicted success and the fast growing rate of these platforms shortly opened new fascinating academic problems; e.g. it is possible to study OSN with tools provided by the SNA science [19]? Is the behavior of OSN users comparable with the one showed by actors of real-life social networks [12]? What are the topological characteristics of OSN [2]? And what about their structure and evolution[16]? Today we exploit computational resources to analyze data

---

[4] http://www.facebook.com

acquired from OSN, trying to answer to these problems. In this work, we analyze connections among almost a million of Facebook users, data collected through some developed *ad hoc* Information Extraction techniques.

This paper is organized as follows: in Section 2 we consider related works on social networks, OSN, etc., in particular regarding data mining experiments and SNA; Section 3 covers aspects of Artificial Intelligence and Information Extraction related to algorithms and techniques used in order to acquire and gather data from Facebook; Section 4 presents collected data, focusing on their statistical analysis, exploiting some tools provided by the SNA science. In Section 5 we try to graphically plot this information, i.e., a large graph where nodes represent users and edges reflect ties among them. Section 6 concludes, providing some suggestions for future work.

2   Related Work

Literature on Web (and social Web) data extraction is growing: Ferrara et al. [10] provided a comprehensive survey on applications and techniques. In [9], Ferrara and Baumgartner developed some techniques for automatic wrapper adaptation. A slightly modified version of that algorithm, relying on analyzing structural similarities inside the DOM tree structure of Facebook friend-list pages, is the core of the agent used here to gather data.

A common SNA task is to discover, if existing, aggregations and subsets of nodes playing similar roles or occupying a particular position in a network [7]. Some strictly connected problems are related to optimizing the visual representation of graphs [4]; for large social networks graphs is not trivial to find a meaningful graphical representation, because of the number of elements to display, and finding algorithms for the planar embedding of the graph, so as reducing (or eliminating) intersecting edges and improving aesthetic and functional characteristics of the graph itself, is part of the solution [5].

Several SNA tools have been developed during the last years: GUESS [1] focuses on improving the interactive exploration of graphs; NodeXL [21], developed as an add-in to the Microsoft Excel 2007 spreadsheet software, provides tools for network overview, discovery and exploration. LogAnalysis [8] helps forensic analysts in visual statistical analysis of mobile phone traffic networks. Jung [17] and Prefuse [14] provide Java APIs implementing algorithms and methods for building applications for graphical visualization and SNA for graphs.

3   Data Extraction and Data Cleaning Aspects

The very first step of a SNA experiment is acquiring data: for this purpose we designed and developed a custom agent, an automaton simulating the behavior of real users, visiting Facebook publicly accessible profiles and automatically extracting relationships among them. Once acquired, information must be collected in some kind of well-structured format; completed this process, data must

be cleaned, removing duplicates and irrelevant information, then they are ready to be used for their purpose.

In order to acquire information about friendship relations, we developed an agent that automatically visits the friend-list page of a real user seed profile, and then recursively, acquires friendship relations visiting friend-list pages of friends of the seed, and so on, down to the third sub-level of friendship relations. Only friendship relations among real users have been acquired, fan pages and companies having been discarded (Facebook provides this filter). This agent acquires information only from profiles in friendship relation with the seed and from publicly accessible profiles, thus respecting the Facebook privacy policies.

We thus obtained an undirected graph composed of 547,302 vertices and 836,468 edges; for privacy reasons only user IDs were collected. The agent was developed in Java, and it embeds a Firefox browser interfaced through XPCOM [5] and XULRunner [6].

Facebook profiles are saved as GraphML [6] nodes with one attribute, namely the Facebook ID. Friendship relations are saved as undirected edges connecting two nodes. Because of the intrinsic nature of the data mining process, it could happen to save parallel edges and multiple instances of the same node. We developed a fast algorithm of data cleaning, running in $O(n \log n)$, exploiting the hash property of the Java HashSet, which, first of all removes all duplicate nodes and, then, fixes all edges in order to link the unique instance of source and target nodes and, finally, deletes the parallel ones.

4 Data Analysis

SNA provides some useful techniques to analyze dynamics of relationships: it is possible to identify models and flows, in the structure of the network, e.g. trying to understand which role actors play in the environment they are placed in. Several statistical algorithms are helpful in discovering key nodes, creating groups of social cohesion [24]. However, it is not trivial to discover models or anomalies while dealing with huge amount of data; in these cases, visualization of information, on the one hand, could be useful to simplify the work of analysis, but on the other hand, could be tricky for several reasons. The computational cost is very high and increases with the dimension of data; it becomes harder and harder to understand graphs showing thousands of nodes and edges, also because of overlapping elements. For these reasons we shall analyze data using filters and clustering methods.

4.1 Metrics and Measures

The following measures for SNA have been standardized by Perer and Shneiderman [20]: overall network metrics (number of nodes, number of edges, density,

---

[5] https://developer.mozilla.org/en/xpcom
[6] https://developer.mozilla.org/en/XULRunner

diameter), node rankings (degree, betweenness and closeness centrality), edge rankings (weight, betweenness centrality), edge rankings in pairs and cohesive subgroups. A short summary of some metrics, evaluated using NodeXL, follows.

Graph Type:	Undirected
Vertices:	547,302
Unique Edges:	836,468
Edges With Duplicates:	0
Total Edges:	836,468
Self-Loops:	0
Connected Components:	2
Single-Vertex Connected Components:	0
Maximum Vertices in a Connected Component:	546,733
Maximum Edges in a Connected Component:	835.9
Maximum Geodesic Distance (Diameter):	10
Average Geodesic Distance:	5.00

Table 1. Overall Network Metrics

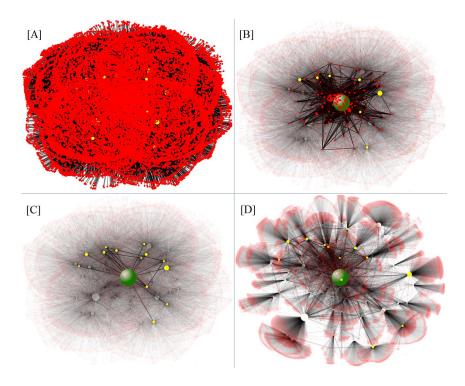|	Minimum	Maximum	Average	Median
Degree	1	4958	3.057	1.000
PageRank	0.269	2120,268	1.000	0.491
Clustering Coefficient	0.000	1.000	0.053	0.000
Eigenvector Centrality	0.000	0.003	0.000	0.000

Table 2. Miscellaneous Metrics

5 Visualization

Analyzing large graphs is not a trivial problem: the computational cost of visualization algorithms, e.g. Fruchterman-Reingold [11], Harel-Koren [13], etc., is critical, and finding useful information is hard. For this reason the SNA relies on filtering data calculating metrics and displaying only relevant information.

We produced several subgraphs for SNA and visualization purposes (see Figure 1). In Figure 1B nodes are arranged according to betweenness centrality (BC). As the definition of BC implies, nodes in the central region of the plot show higher values of BC, thus occurring in a correspondingly higher number of shortest paths connecting all the nodes each others. Intuitively, nodes with higher BC values have higher relevance as they represent a potential efficient way of establishing friendship relations among peripheral nodes.

6 Conclusions

In this preliminary work we focused on the possibility of extracting relevant information about relationships from Facebook. We developed an automaton

Fig. 1. [A] Visualization of a 25,000 nodes subgraph; [B] Top 50 nodes ordered by betweenness centrality; [C] Nodes with high betweenness centrality (greater than 10 millions); [D] Clusterization after 10 iterations of the Fruchterman-Reingold algorithm.

for data mining and cleaning, gathered a sample dataset of almost a million of connections, and finally analyzed data applying SNA tools and techniques. Our purpose is to continue acquiring data and we already developed a more efficient way of data mining, and then to improve algorithms for data analysis and visualization, e.g. exploiting the auto parallelization and High Performance Computing techniques to handle in the most efficient way the huge amount of useful information we can gather from Facebook.

References

1. Adar, E.: Guess: a language and interface for graph exploration. In: CHI '06: Proceedings of the SIGCHI conference. pp. 791–800. ACM, NY, USA (2006)
2. Ahn, Y.Y., Han, S., Kwak, H., Eom, Y.H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. Proceedings of the 16th international conference on World Wide Web pp. 853–844 (2007)
3. Barabasi, A.L., Crandall, A.R.E., Reviewer: Linked: The new science of networks. American Journal of Physics 71(4), 409–410 (2003)
4. Battista, G.D., Eades, P., Tamassia, R., Tollis, I.: Algorithms for drawing graphs: an annotated bibliography. Computational Geometry 4(5), 235–282 (Oct 1994)

5. Boyer, J.M., Myrvold, W.J.: On the Cutting Edge: Simplified O(n) Planarity by Edge Addition. J. of Graph Algorithms and Applications 8(3), 241–273 (2004)
6. Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., Marshall, M.S.: GraphML Progress Report Structural Layer Proposal. Lecture Notes in Computer Science 2265, 109–112 (2002)
7. Carrington, P.J., Scott, J., Wasserman, S.: Models and methods in social network analysis. Cambridge ; New York : Cambridge University Press (2005)
8. Catanese, S., Fiumara, G.: A visual tool for forensic analysis of mobile phone traffic. In: MiFor '10: Proceedings of the Second ACM workshop on Multimedia in forensics (to appear) (2010)
9. Ferrara, E., Baumgartner, R.: Automatic Wrapper Adaptation by Tree Edit Distance Matching. In: ICTAI '10: 22th IEEE International Conference on Tools with Artificial Intelligence (to appear) (2010)
10. Ferrara, E., Fiumara, G., Baumgartner, R.: Web Data Extraction, Applications and Techniques: A Survey. Technical Report (2010)
11. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. Software: Practice and experience 21(11), 1129–1164 (1991)
12. Garton, L., Haythornthwaite, C., Wellman, B.: Studying online social networks. Journal of Computer-Mediated Communications 3(1) (1997)
13. Harel, D., Koren, Y.: A fast multi-scale method for drawing large graphs. In: AVI '00: Proceedings of the working conference on Advanced visual interfaces. pp. 282–285. ACM, New York, NY, USA (2000)
14. Heer, J., Card, S.K., Landay, J.A.: prefuse: a toolkit for interactive information visualization. Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05 p. 421 (2005)
15. Kleinberg, J.: The small-world phenomenon: an algorithm perspective. In: STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing. pp. 163–170. ACM, New York, NY, USA (2000)
16. Kumar, R., Novak, J., Tomkins, A.: Structure and Evolution of Online Social Networks. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 611–617 (2006)
17. Madadhain, J., Fisher, D., Smyth, P., White, S., Boey, Y.: Analysis and visualization of network data using JUNG. Journal of Statistical Software VV(II) (2005)
18. Milgram, S.: The Small World Problem. Psychology Today 1(1), 61–67 (1967)
19. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and Analysis of Online Social Networks. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement pp. 29–42 (2007)
20. Perer, A., Shneiderman, B.: Balancing systematic and flexible exploration of social networks. IEEE TVCG 12(5), 693–700 (2006)
21. Smith, M.a., Shneiderman, B., Milic-Frayling, N., Mendes Rodrigues, E., Barash, V., Dunne, C., Capone, T., Perer, A., Gleave, E.: Analyzing (social media) networks with NodeXL. Proceedings of the fourth international conference on Communities and technologies - C&T '09 p. 255 (2009)
22. Staab, S., Domingos, P., Mika, P., Golbeck, J., Ding, L., Finin, T., Joshi, A., Nowak, A., Vallacher, R.R.: Social networks applied. IEEE Intelligent Systems 20(1), 80–93 (2005)
23. Travers, J., Milgram, S.: An experimental study of the small world problem. Sociometry 32(4), 425–443 (1969)
24. Wasserman, S., Faust, K.: Social network analysis: methods and applications. Cambridge University Press, 1st edn. (1994)