Some Issues related to the Mining of OSNs represented as Graphs

David F. Nettleton[1,2]

[1]Web Research Group, Departament of Information Technology and Communications, Universitat Pompeu Fabra, Barcelona.

[2]IIIA-CSIC, Bellaterra.

19th February 2013

Summary

- This brief talk will consider some of the issues which graph data miners may encounter when analyzing Online Social Networks represented as graphs.

- Such issues include the elicitation of a community structure, finding similar sub-graphs and computational cost issues, among others.

Issues

- We will briefly look at the following issues:
  - The representation of an OSN as a graph
  - Elicitation of a community structure
  - Finding similar sub-graphs
  - Computational cost issues

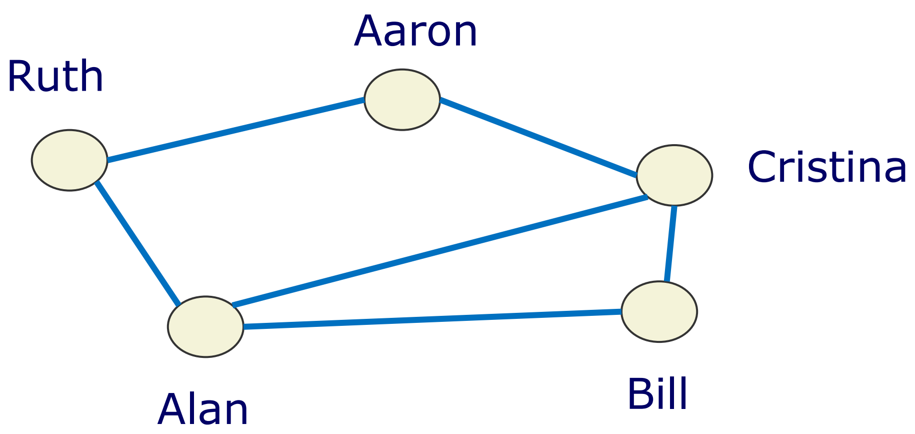
Representation of an OSN as a graph

- A graph is comprised of nodes and edges, but its easy to misrepresent an OSN: [1]

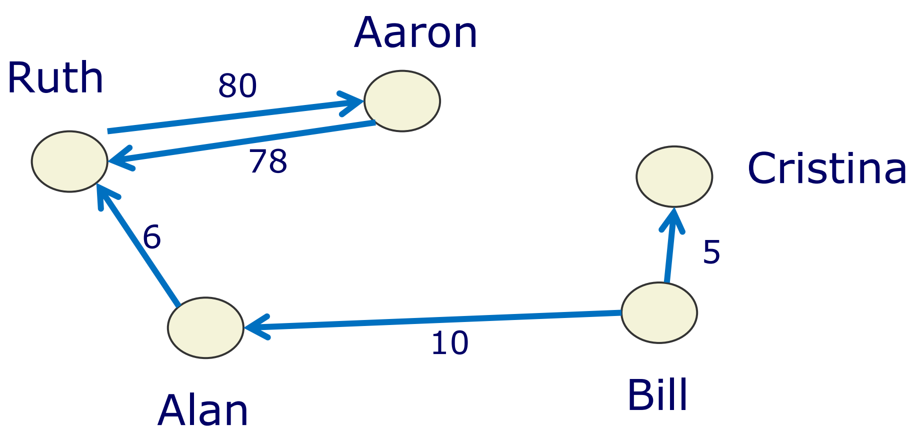
  - What type of activity between nodes is chosen to define a link?
    - Some key data may be unavailable.
  - Related to the first point, what is the minimum activity level (by frequency or latency) in order for a link to appear between two nodes?
  - What information is available about each node individually and the nature of the graph as a whole.
  - What does the 'user' wish to DO with the graph once the OSN is represented?

Representation of an OSN as a graph

Existence of an edge implies that have mutually accepted friendship request in OSN application. No weights on edges.



Existence of an edge implies at least 5 messages sent/received over last 3 months. Weights on edges indicate number of messages sent/received.



Elicitation of a community structure

Algorithm 1: Newman's algorithm [2]

- Extracts the communities by successively dividing the graph into components, using Freeman's between-ness centrality measure until modularity Q is maximized.

- Modularity (Q): Is the measure used to quantify the quality of the community partitions 'on the fly'. Usual range: [0.3 - 0.7].

- Problem: it's slow

Extraction of a community structure

Algorithm 2: Blondel's 'Louvain' method [3]

1. The method looks for smaller communities by optimizing modularity locally.

2. Then it aggregates nodes of the same community and builds a new network whose nodes are communities.

Steps 1 and 2 are repeated until modularity Q is maximized.

- This algorithm is used in the Gephi graph processing software.
- It's significantly faster than Newman's method, because, due to the aggregation in step 2, after each iteration, there are progressively less nodes to process.

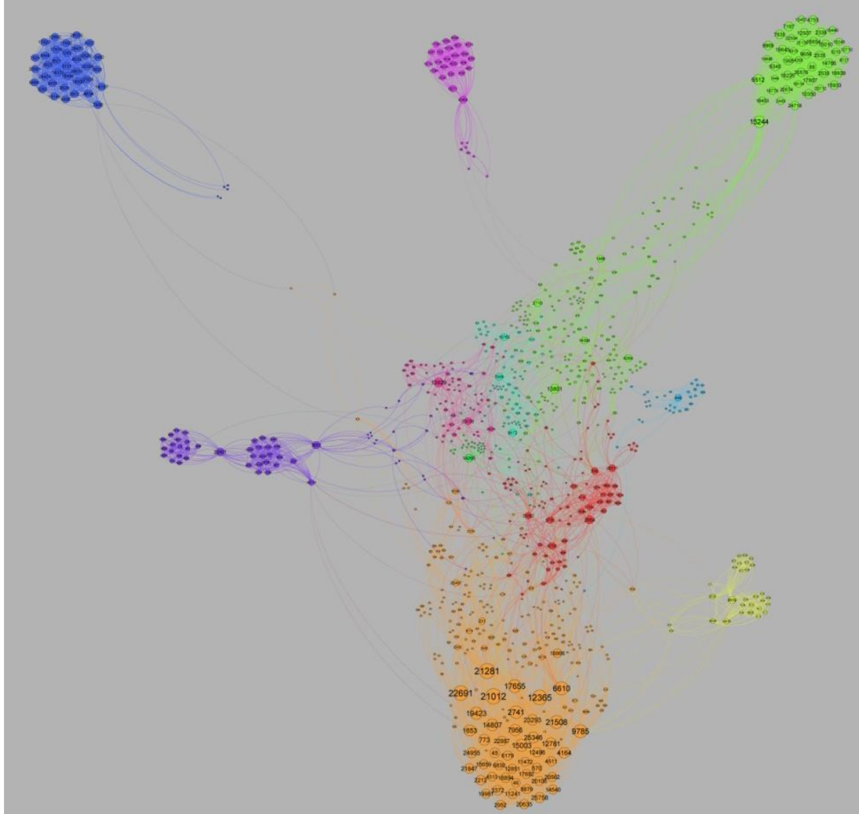
Extraction of a community structure

Problems with results of a community extraction [1,4]

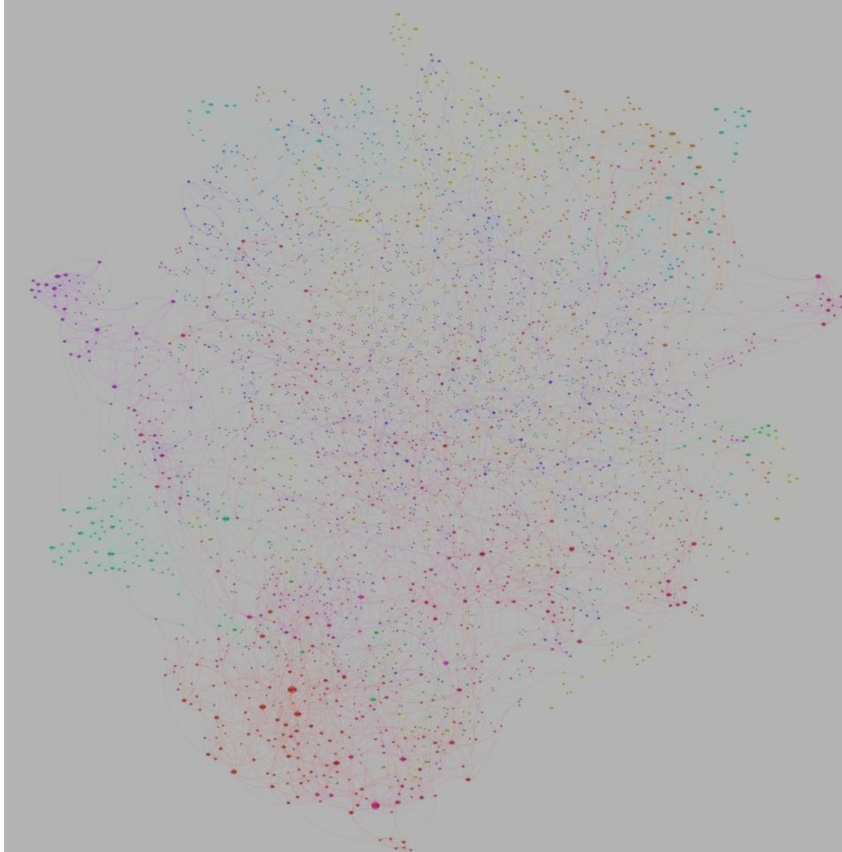
1. Process is stochastic. May produce slightly different community structure each time.
2. Interpreting the communities
3. Identifying key nodes, frontiers
4. Quality of resulting structure.

Extraction of a community structure

Dataset: arXiv-GrQc [5]



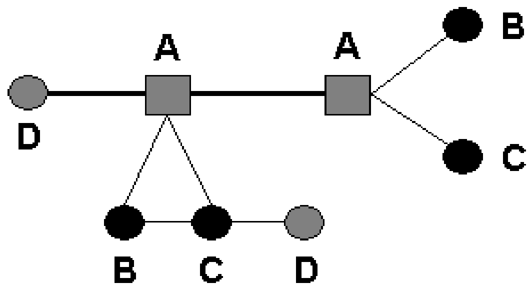
Dataset: Facebook New Orleans [6]

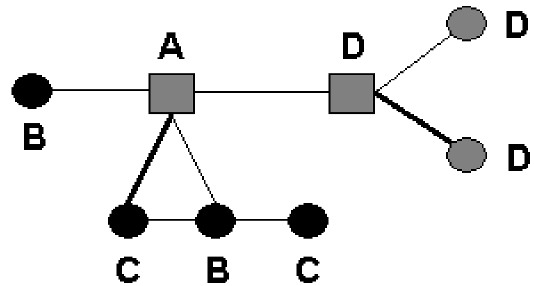


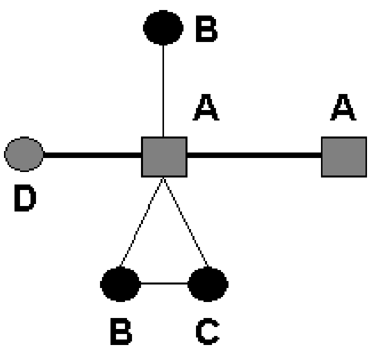
Finding similar sub-graphs

1. The most powerful tool for finding exact sub-graphs is an isomorphism matcher
    1. The VF2 algorithm [7] has become an 'industry standard' for isomorphism matching
    2. Isomorphism matching is more important for some domains, such as chemical and pharmaceutical analysis.
2. But maybe we don't need an exact match on topological properties. Perhaps, for our needs, we just want an approximation based on the node/edge characteristics [1,4]
    1. Type of node
    2. Volume of traffic between edges
    3. Characteristics of one or more neighbour nodes

Finding similar sub-graphs







Which two graphs are most similar?

Computational cost issues [1]

1. One of the key problems of processing large graphs is the NP-completeness of many of the typical processes
    1. Isomorphism matching, average path length
    2. Entropy based approaches
2. The first measure is to use an efficient representation of the graph, depending on its characteristics:
    1. Adjacency list /matrix for nodes and connexions
    2. Storage of sparse data, Hash tables, …
3. Processing:
    1. Often, a good approximation is sufficient, without having to exhaustively process the whole graph.
    2. Sampling, streaming for very large graphs
    3. Hardware (especially Ram memory) is important

References

[1] D. F. Nettleton, Data Mining of Social Networks Represented as Graphs. In Press, Computer Science Review (2013), doi:10.1016/j.cosrev.2012.12.001

[2] M.E.J. Newman, M. Girvan, Finding and Evaluating Community Structure in Networks, Phys. Rev. E 69, 026113, 2004.

[3] V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefebure. Fast Unfolding of Communities in Large Networks, in Journal of Statistical Mechanics: Theory and Experimentation (10), 2008, pp. 1000.

[4] N. Martínez Arqué, D. F. Nettleton, Analysis of On-line Social Networks Represented as Graphs – Extraction of an Approximation of Community Structure Using Sampling, Proc. Modeling Decisions for Artificial Intelligence (MDAI) 2012, Girona, Catalunya. Lecture Notes in Artificial Intelligence (LNAI), Vol. 7647, pp. 149-160 (2012)

[5] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, 2009. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. Internet Mathematics Vol. 6, No. 1: 29–123.

[6] B. Viswanath, A. Mislove, M. Cha, K.P. Gummadi, 2009. On the Evolution of User Interaction in Facebook. In Proc. 2nd ACM workshop on Online social networks WOSN'09, Barcelona, Spain, pages 37-42.

[7] L.P. Cordella, P. Foggia, C. Sansone, M. Vento, An Improved Algorithm for Matching Large Graphs, in Proc. 3rd IAPR-TC-15 International Workshop on Graph based Representations, Cuen, Italy, 2001, pp. 149-159.

Thank you for your attention !