Comparative study of community detection algorithms on social networks

Andreas Kalaitzakis, Harris Papadakis, Paraskevi Fragopoulou

Department of Applied Informatics & Multimedia

Technological Educational Institute of Crete

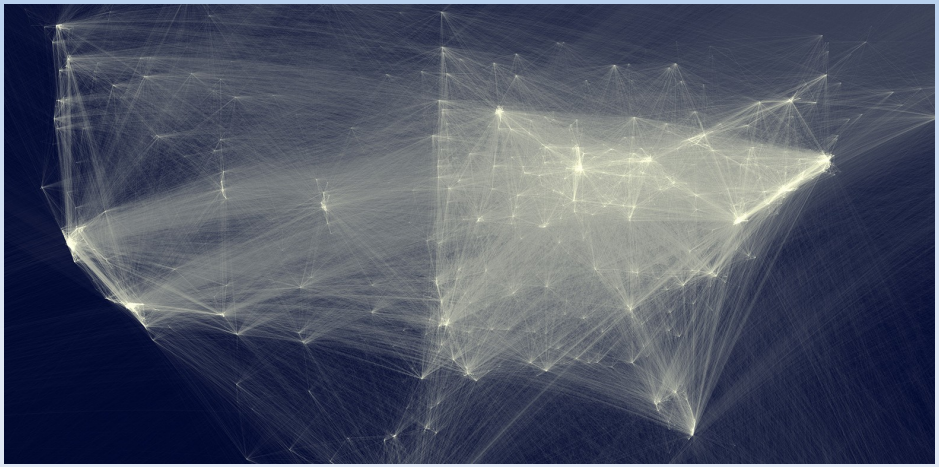


Outline

- Community Detection Problem
- Algorithms
  - Newman's Algorithm
  - CiBC
  - Fortunato's Algorithm
  - Bridge Bounding
  - Clique Percolation
- Experimental Evaluation on Synthetic Data
- Applying Community Detection on Real-World Networks
  - Analyzing Underlying Friendship Graph
  - Analyzing Population and User Activity Evolution
- Conclusions

Community Detection I

- Given a network
  - Determine the existance of any natural division
  - Detect the surrounding community of a vertex
  - Reveal its community structure



Community Detection II

- Community Detection is not new to literature
- It has been been applied to a wide range of well established scientific disciplines:
  - Discrete Mathematics
  - Biology
  - Social Sciences
  - Criminology
  - Physics
  - Computer Science
- New parameters that chaperon the problem
  - Advent of Web 2.0 technology
  - Appearance of the first social networks

Community Detection Algorithms Classification I

- Local Perspective
  - *Clique Percolation Method*
  - *Fortunato's Algorithm*
  - *Bridge Bounding*
- Global Perspective
  - *Newman's Algorithm*
  - *CiBC*

- Algorithms adopting local perspective
  - Are based on local topological properties
  - Do not require the whole network
  - Are suitable for real-world networks manipulation

Community Detection Algorithms Classification II

- Supporting Overlapping Community Structure
  - *Clique Percolation Method*
  - *Bridge Bounding Method*
  - *Fortunato's Algorithm*
- Supporting Distinct Community Structure
  - *Newman's Algorithm*
  - *CiBC*

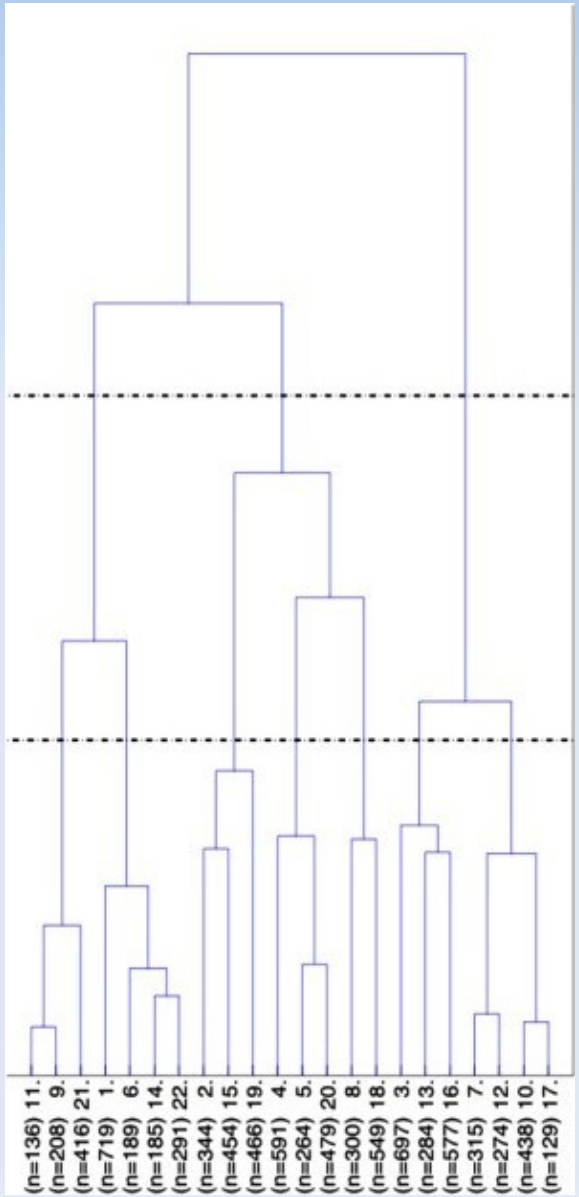
- Various definitions exist regarding what community stands for
  - *Strong community*
  - *Weak community*
- Supporting overlapping community structure lies on the adopted definition

Newman's Algorithm I

- Is based on the betweeness centrality measure
  - An edge that belongs to many shortest paths between nodes has **high betweeness value** and is more likely to be an inter-community edge
- Follows a divisive approach
- *General Form of the Algorithm*
  - *Calculate betweeness scores for all edges of the network*
  - *Remove the edge with the **highest betweeness score***
  - *Recalculate betweenness for all remaining edges*
  - *Repeat from step 2*

Newman's Algorithm II

- *prō*
  - Robust
- *contrā*
  - Unable to manipulate evolving graphs
  - Computationally intensive
  - Requires whole network



CiBC Method I

- Designed to identify web communities
- Based on betweeness centrality measure
- Identifies overlapping community structure

- *General Form of the Algorithm*
  - *Calculate betweeness scores for all nodes and sort them by ascending score*
  - *Create cliques starting with the nodes with the lowest betweeness score*
  - *Merge appropriately the cliques that were created in phase 2*

CiBC Method II

- *prō*
  - Tolerable computational cost
- *contrā*
  - Unable to manipulate evolving graphs
  - Uses global functions
  - Limited success

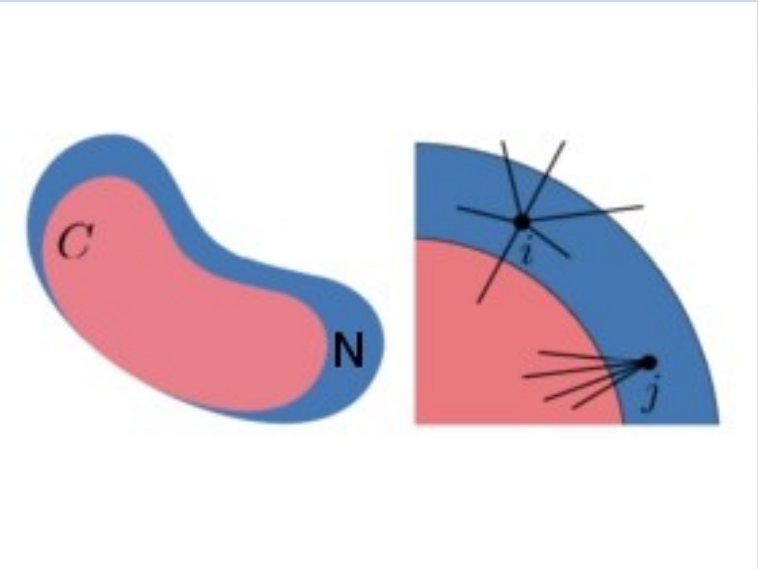
Fortunato's Algorithm I

- Based on a **“Fitness”** measure
  - Corresponds to the correlation that exists between a node and a community
- Operates using a flooding technique
- Designed to identify overlapping community structure

- *General Form of the Algorithm*
  - *Initiate a subgraph G including node A*
  - *Perform a loop over all neighboring nodes of G not included in G*
  - *The neighbor with the largest fitness is added to G , yielding a larger subgraph G'*
  - *The fitness of each node of G' is recalculated*
  - *If a node turns out to have negative fitness, it is removed from G' , yielding a new subgraph G''*
  - *If 5 occurs, repeat from 4, otherwise repeat from 2 for subgraph G*

Fortunato's Algorithm II

- *prō*
  - Does not require the whole network
  - Low computational cost
  - Highly successful
- *contrā*
  - Unable to manipulate evolving graphs



Bridge Bounding Method I

- Is based on the local clustering coefficient measure
  - Acts as an indication of how close its neighbors are to form a complete subgraph
- Operates identifying edges that act as community boundaries

- *General Form of the Algorithm*
  - *Initiate a subgraph G including node A*
  - *Progressively expand G through a flooding process*
  - *Terminate expansion when edges that act as boundaries are found*
  - *Repeat procedure until total community structure is revealed*

Bridge Bounding Method II

- *prō*
  - Does not require the whole network
  - Low computational cost

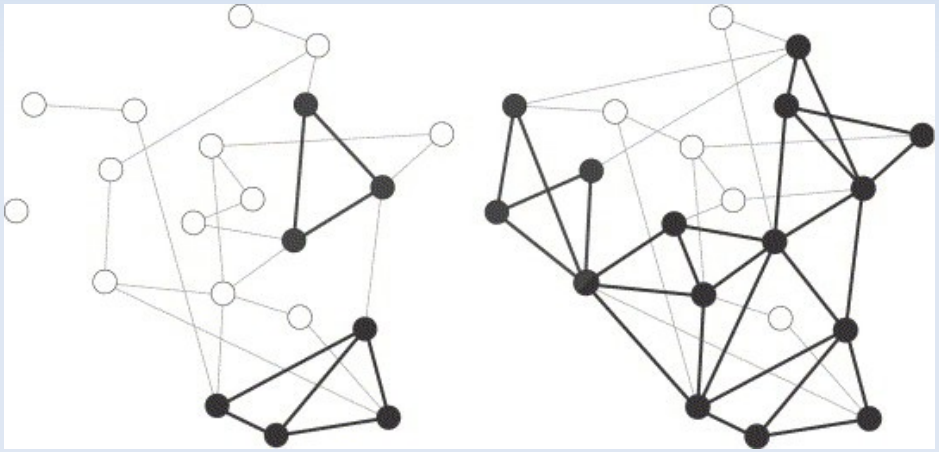
- *contrā*
  - Unable to manipulate evolving graphs
  - Very limited success

Clique Percolation Method I

- Bases its operation on **“K-Qlique”** structures
  - Corresponds to a maximally connected subset of nodes. Each node must be connected to every other node that participates into the clique.
- Two K-cliques are considered to be adjacent if they share K-1 nodes
- A community is defined as the maximal union of K-cliques that can be reached from each other through a series of adjacent K-cliques

Clique Percolation Method II

- *prō*
  - Does not require the whole network
  - Low computational cost
- *contrā*
  - Unable to manipulate evolving graphs
  - Limited success



Experimental Evaluation on Synthetic Networks I

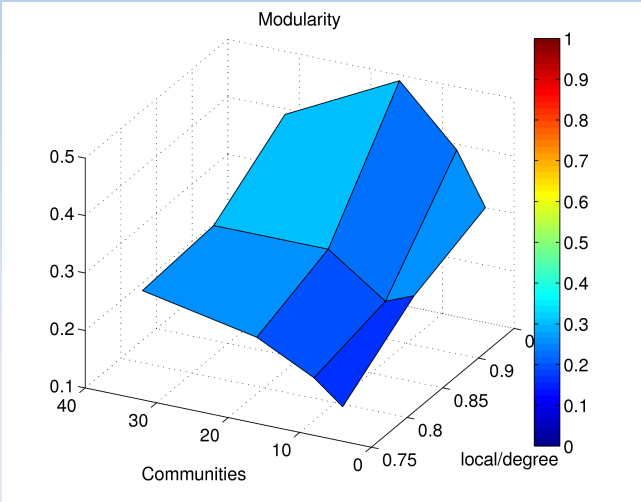
- Was based on a series of benchmark graphs with known structure
- Parameters
  - *Number of graph nodes (N)*
  - *Number of communities (Comm)*
  - *Node degree (degree)*
  - *Ratio of intra-community edges to node degree (local/degree)*

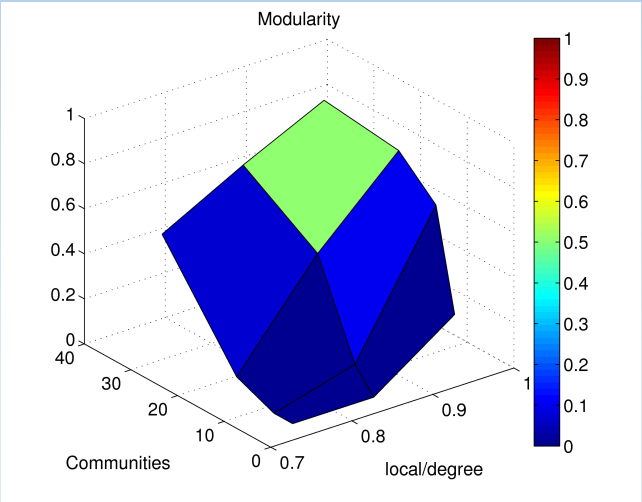
N	512, 1024
Comm	4, 8, 16, 32
Degree	10, 20, 30
Local/ Degree	0.75, 0.85, 0.95

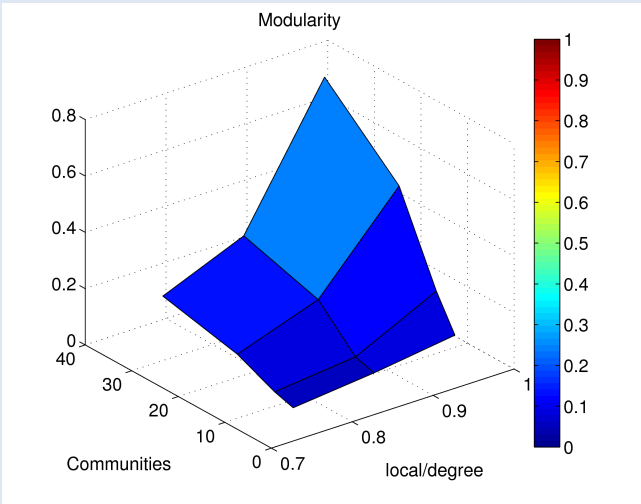
Experimental Evaluation on Synthetic Networks II

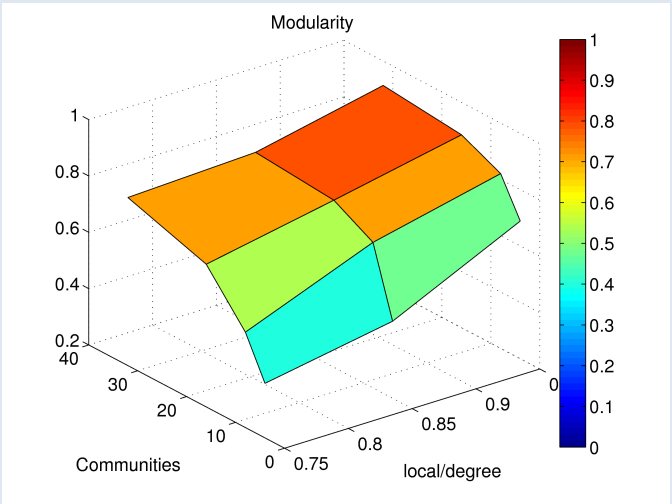
- Algorithms' output quality is evaluated by the well established Newman's **“Modularity”** measure
  - Indicates the extent to which a given community partition is characterized by high number of intra-community edges compared to inter-community ones

Experimental Evaluation on Synthetic Networks III







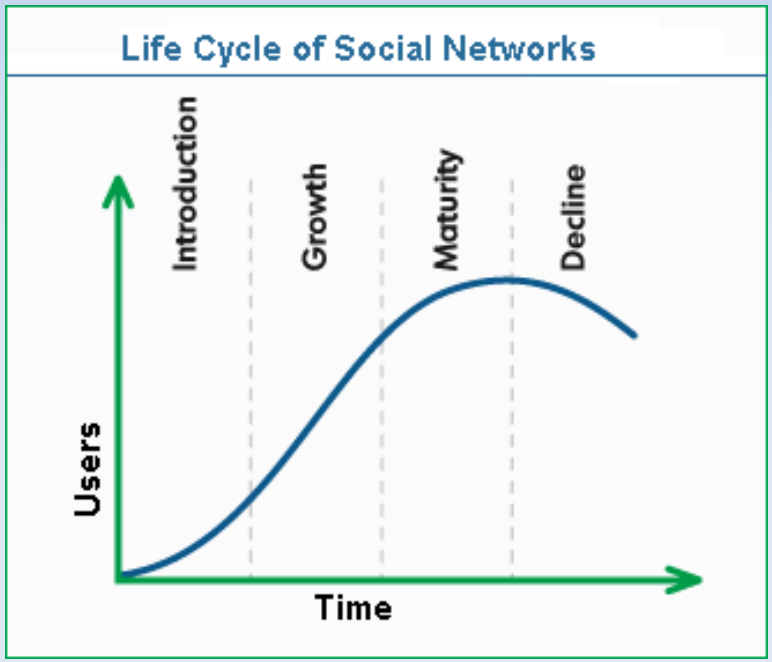


Applying Community Detection on Real-World Networks I

- Can algorithms be applied equally successfully to real-world networks?
  - *If not, what are the properties that affect their operation?*
- MySpace Social Network
  - *Launched in August 2003*
  - *Sold in July 2005 for **$580** million*
  - *From 2005 until early 2008, MySpace **is the most visited social networking site** in the world*
  - *Since then, the number of Myspace users has declined steadily in spite of several redesigns*
  - *In June 2011, Specific Media Group and Justin Timberlake jointly purchased the company for approximately **$35** million*
  - *As of June 2012, Myspace was ranked **161st** by total web traffic*

Modern Social Networks I

- *Suggesting that Social Networks are characterized by a Cycle of Life..*
- *We descry four different stages*
  - *Introduction*
  - *Growth Stage*
  - *Maturity*
  - *Decline*



Modern Social Networks II

- Introduction Stage
  - The network is discovered and new users begin to register to the service
  - Early adopters get their friends to join
- Growth Stage
  - Number of registrations is reduced compared to the initial stage
  - In this phase there is still an increase to new registrations
- Maturity Stage
  - The rate of new users registrations is stabilized
  - Providing new features to the existing service can temporarily reduce the user turnover rate
- Decline Stage
  - More people are quitting the network than joining it
  - Users immigrate to new, more popular social networks

Data Mining & Methodology I

- Data Mining from Social Networks is a non-trivial task
  - Social networks tend to hide informations from public view
  - Such policies are adopted in an effort to prevent more users from abandoning the network at the stage of decline
- MySpace Network favors network analysis in several ways
  - Unlike the plurality of today's social networks, user profiles are accessible via HTTP requests
  - This way we do not need to rely on the availability of periodically released enormous static datasets
  - MySpace allocates numeric user IDs in a sequential pattern
- Using html parsing to preprocess the downloaded profiles we extracted all necessary attributes for our research

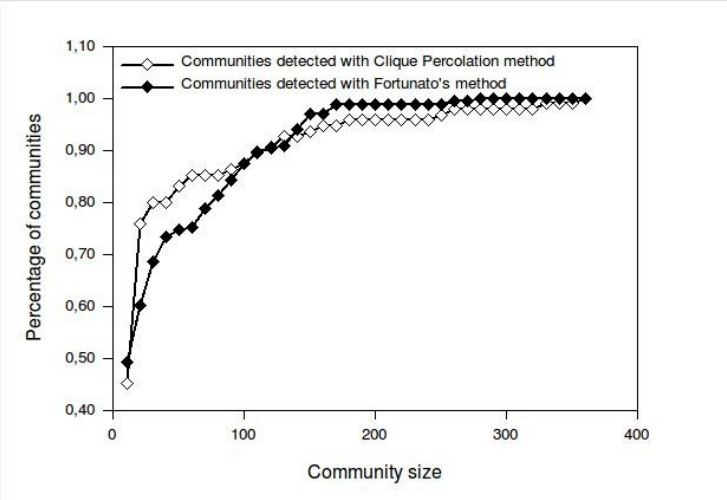
Data Mining & Methodology II

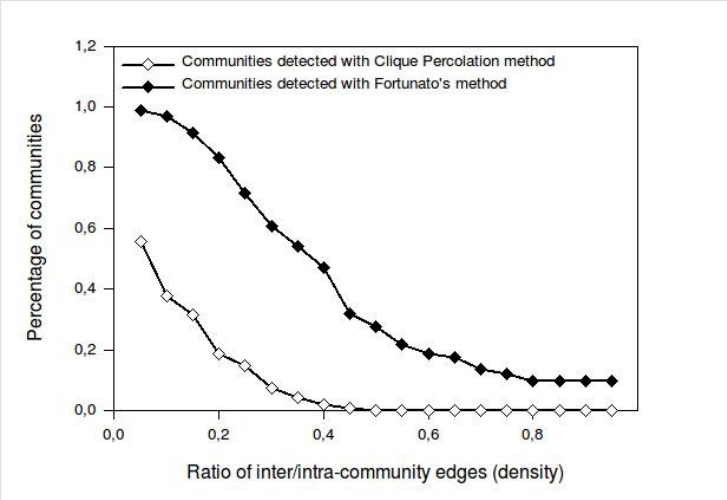
- Depending profile's status a number of information regarding the user was collected including:
  - *User ID*
  - *Status attribute = {"Private", "Public", "Invalid"}*
  - *Alias*
  - *Number of friends*
  - *Number of views*
  - *Member since date*
  - *Last login date*
- Out of 999.937 randomly collected profiles:
  - 403183 were found to be public
  - 177664 were found to be private
  - 419090 were found to be invalid

Analyzing Underlying Friendship Graph I

- What type of communities do nodes form?
  - Are users that participate in communities more active than users that are not?
  - Do inactive users tend to abandon MySpace individually or the same tendency exists for their surrounding community?
- In order to answer to these questions we proceeded with applying Fortunato's and Qlique Percolation methods
  - Using Fortunato's method we successfully detected:
    - 171 communities
    - Containing 6132 members in total / ~37 members per detected community
    - 18259 of fetched profiles during the operation
  - Using Clique Percolation method we successfully detected:
    - 201 communities
    - Containing 3576 members in total / ~18 members per detected community
    - 176577 of fetched profiles during the operation

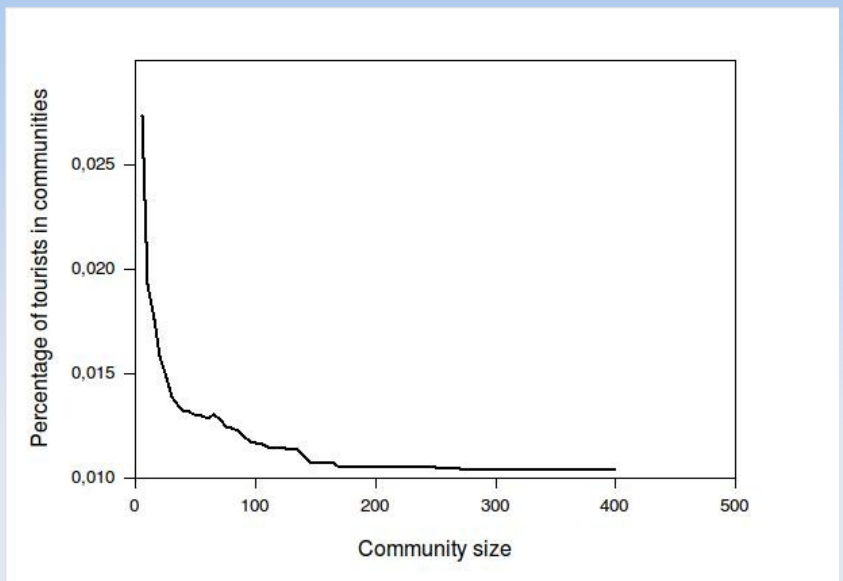
Analyzing Underlying Friendship Graph II





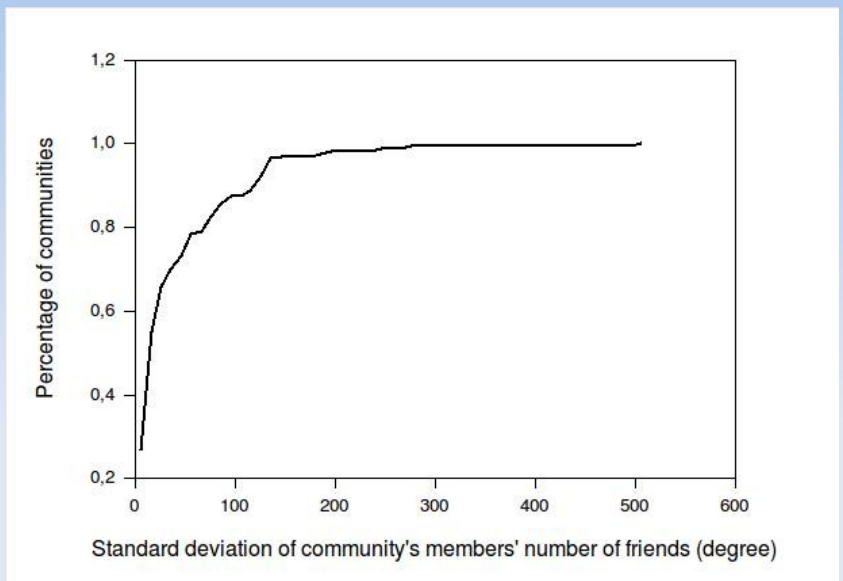
- Although Clique Percolation proved to be extremely fast in detecting the surrounding community of randomly selected nodes it was no match to Fortunato's method in terms of output's quality
  - Fortunato succeeded better results in terms of community density and community size
  - Moreover Fortunato's algoritmh required significantly less information during its operation
- Taking into account the previously mentioned differences in terms of performance we continued our analysis based on Fortunato's results

Analyzing Underlying Friendship Graph III



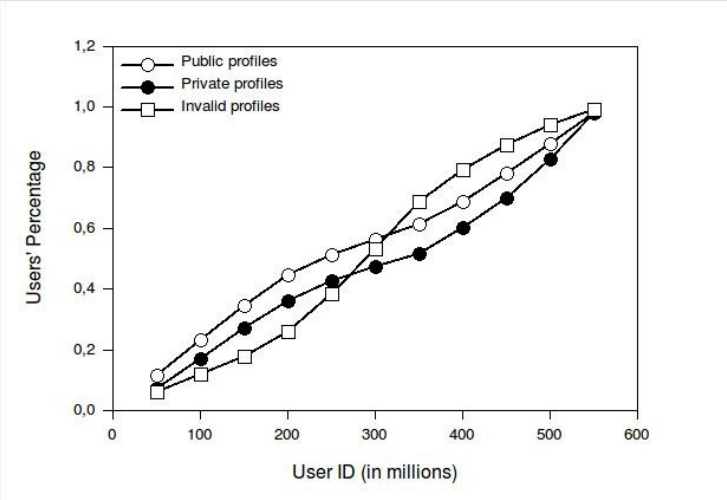
- We focused our efforts into locating members that had a very short stay in the network
  - More than 85% of the communities consist of almost 90% of users that showed their presence in the network at least once in the last one hundred days
  - Moreover, we observed an extremely low percentage of participation of tourists into community structures
- Although we have indications that the participation of a user in a community, inhibits him to abandon MySpace

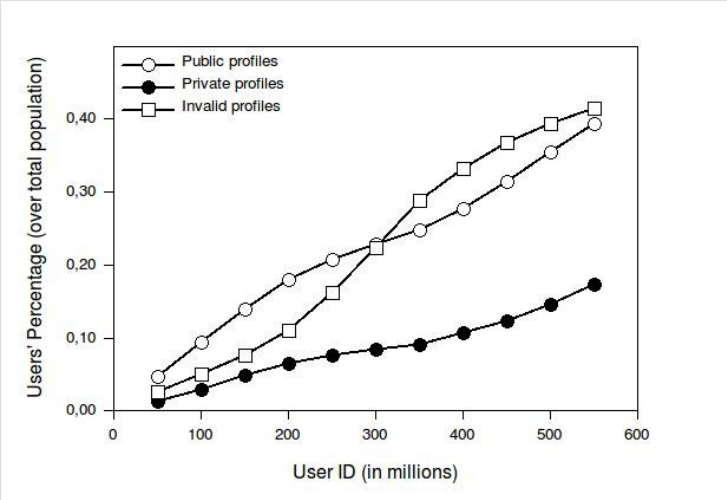
Analyzing Underlying Friendship Graph IV



- Interestingly, the members of the detected communities show a significantly low dispersion over the total degree
  - Moreover 60% of the communities have a standard deviation lower than 20
  - 80% of the communities have a standard deviation lower than 50

Analyzing Population Evolution



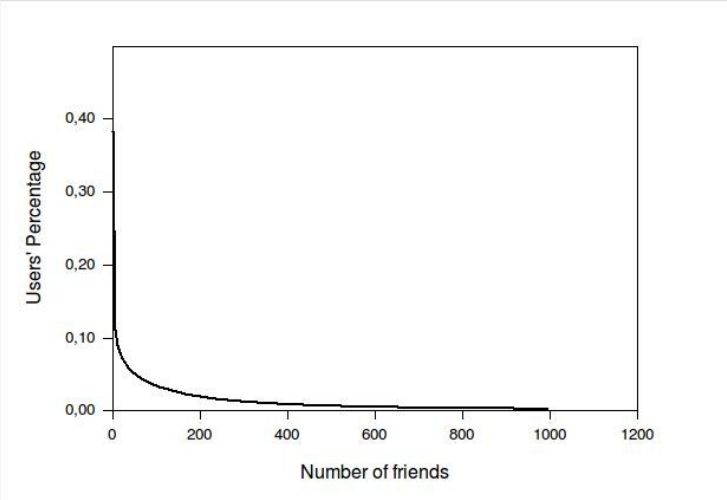


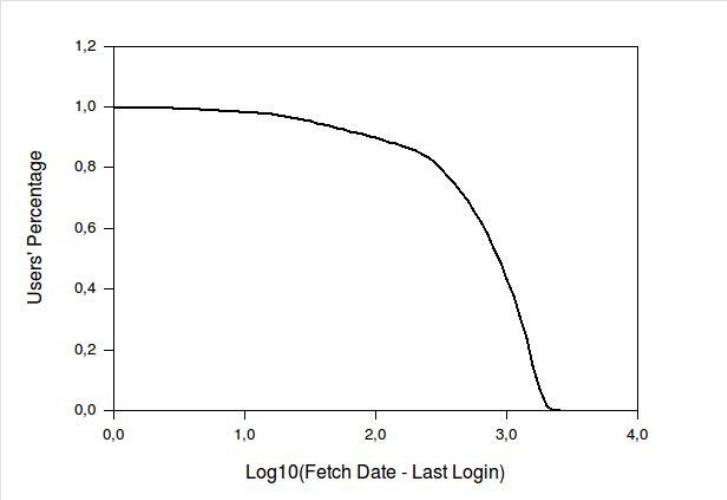
- For all recorded IDs we followed the change in the synthesis of the population based on the percentage of profile status over time
  - For a large initial portion of the ID space, the percentage of public profiles is higher than that of the private profiles
  - Invalid profiles **overpassed** Private ones, and a little later they also **exceeded** Public profiles

Analyzing User Activity I

- We approached the term activity in three different ways
  - Number of friends
  - Number of profile views
  - How many days ago did the user log in for the last time?
- Users characterized as **“tourists”**
  - The term refers to users that had a very short stay in the network
  - The interval expressed in days, for a user to be characterized as “tourist” was defined as the period of ten days

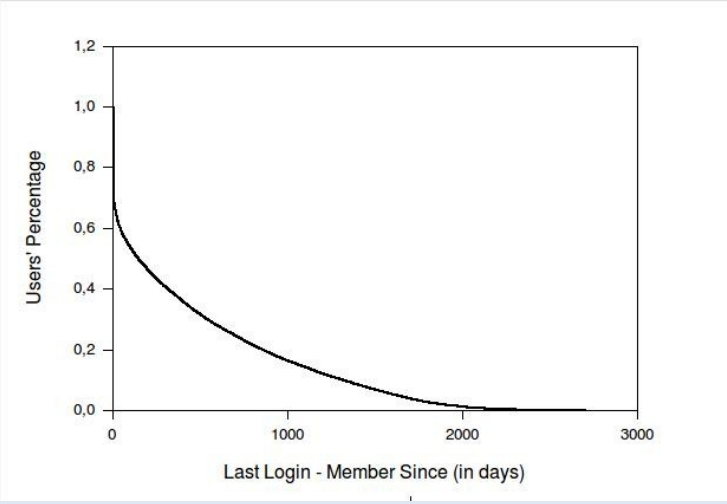
Analyzing User Activity II

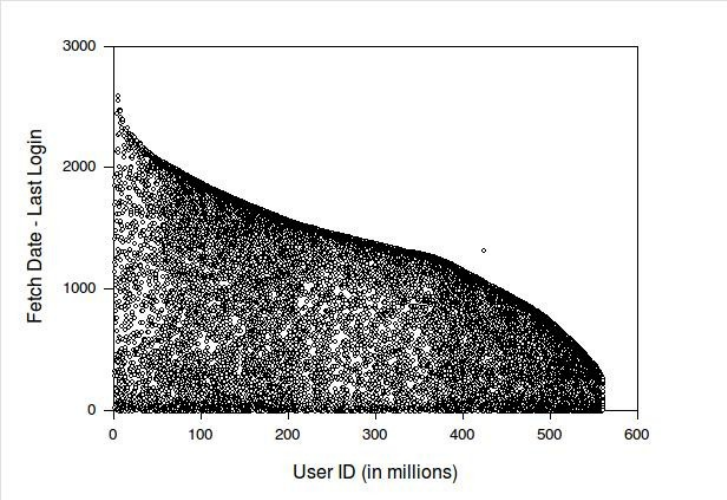




- The plurality of MySpace users has a relatively small Degree
  - Figure shows that only about 5% of Myspace's population has more than 100 friends
- According to our statistics 9 out of 10 users have not logged into the service for the last one hundred days
  - 1.8% of profiles did actually login, up to ten days before the fetch process
  - 8,3% were the users that logged in up to one hundred days before profile fetching took place

Analyzing User Activity III





- An important fraction of Myspace's population had a short stay in the network
  - About 33% percent of the users abandoned the network between one to ten days after signing up
  - Another 13% percent has been found to be active for a period of ten to one hundred days
- As a last indication for user activity, we study the relationship between a user ID and the users Last Login date

Conclusions I

- Algorithms proved to be more or less robust when dealing with the synthetic networks, however they struggled in detecting the surrounding community of a vertex when applied in real-world networks
- Motivated by this observation we tried to conclude into the crucial parameters that leads to this inadequacy examining the underlying structure of MySpace social network
  - The most important observed characteristic of MySpace and hence of all social networks and collaborative environments has to be its tendency to evolve
  - Size-wise community structures proved to be decent allowing algorithms to proceed their operation enhancing their situational awareness
  - Although strong correlation was demonstrated on the number of friends among members of the same community, the revealed communities are characterized by poor density

Conclusions II

- There is still plenty of room for improvement of existing solutions and for the derivation of new ones
  - Static manipulation of data rendered prohibitive demanding more flexible solutions
  - Need to devise community detection algorithms for graphs that tend to evolve over time
- Future Work
  - Presume on the extracted information for the benefit of the design of a new algorithm
  - Examine algorithms performance from an information retrieval point of view

Questions

Questions are guaranteed in life;
*answers are **not***

References I

(1)A. M. Kaplan and M. Haenlein, “Users of the world, unite! The challenges and opportunities of Social Media” Business Horizons, 2010.

(2)B. Buter, N. Dijkshoorn, D. Modolo, Q. Nguyen, S. van Noort, B. van de Poel, A. Ali, and A. Salah1. Explorative visualization and analysis of a social network for arts: The case of deviantart. Journal of Convergence Volume, 2(1), 2011.

(3)G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee. Self-organization and identification of web communities. IEEE Computer, 35:66–71, March 2002.

(4)D. Katsaros, G. Pallis, K. Stamos, A. Vakali, A. Sidiropoulos, and Y. Manolopoulos. Cdns content outsourcing via generalized communities. IEEE Transactions on Knowledge and Data Engineering, 21:137–151, 2009.

(5)M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2):026113, Feb 2004.

(6)M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, 99(12):7821–7826, June 2002.

(7)M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69(2):026113, Feb 2004.

References II

(8) S. Papadopoulos, A. Skusa, A. Vakali, Y. Kompatsiaris, and N. Wagner. Bridge bounding: A local approach for efficient community discovery in complex networks. Technical Report arXiv:0902.0871, Feb 2009.

(9) A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics, 11(3):033015+, March 2009.

(10) I. Derenyi, G. Palla, and T. Vicsek. Clique Percolation in Random Networks.Physical Review Letters, 94(16):160–202, Apr 2005.

(11) B. Buter, N. Dijkshoorn, D. Modolo, Q. Nguyen, S. van Noort, B. van de Poel, A. Ali, and A. Salah1. Explorative visualization and analysis of a social network for arts: The case of deviantart. Journal of Convergence Volume, 2(1), 2011.

(12) MySpace. http://www.myspace.com.

(13) Facebook. http://www.facebook.com.

(14) Y.-Y. Ahn, S. Han, H. Kwak, M. S., and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. May 2007.

(15) W. W. Mojtaba Torkjazi, Reza Rejaie. Hot Today, Gone Tomorrow: On the Migration of MySpace Users. WOSN'09, Barcelona, Spain, pages 43–48, August 17 2009.

References III

(15)J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. August 2005.

(16)R. Kumar, J. Novak, and A. Tomkins. Structure and the Evolution of Online Social Networks. August 2006.

(17)G. Pallis, D. Zeinalipour-Yazti, and M. D. Dikaiakos. Online Social Networks: Status and Trends. New directions in Web Data Management, 1, 2011.

(18)A. Lancichinetti and S. Fortunato. Community detection algorithms: a comparative analysis.Physical Review E, 80(5 Pt 2):056117, Sep 2009.

(19)S. E. Schaeffer. Graph clustering.Computer Science Review, 1(1):27 – 64, 2007.

(20)G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society, June 2005.