Mining Historical Social Data for Detecting Persistent Labeled Communities

Master Thesis of: Evdoxia Papadopoulou
Supervisor: Koloniari Georgia

University of Macedonia
MSc in Computational Methods and Applications
Thessaloniki, July 2018

Content

- Objectives
- Related Work
  - Graph Clustering
  - Time-evolving Graphs
  - Graphs and Labels
- Methodology
  - Communities Detection
  - Labels Extraction
  - Stable Communities Identification
- Case Study
  - Dataset
  - Parameters Tuning
  - Experimental Results
- Conclusion
- Future Work

Content

- Objectives
- Related Work
  - Graph Clustering
  - Time-evolving Graphs
  - Graphs and Labels
- Methodology
  - Communities Detection
  - Labels Extraction
  - Stable Communities Identification
- Case Study
  - Dataset
  - Parameters Tuning
  - Experimental Results
- Conclusion
- Future Work

Objectives

- Study the evolution of time-evolving networks at a community level
- Propose a methodology that can be used to detect stable communities and stable characteristics of communities that persist during time.
- Test the proposed approach in co-authorship network (DBLP)

Content

- Objectives
- Related Work
  - Graph Clustering
  - Time-evolving Graphs
  - Graphs and Labels
- Methodology
  - Communities Detection
  - Labels Extraction
  - Stable Communities Identifications
- Case Study
  - Dataset
  - Parameters Tuning
  - Experimental Results
- Conclusion
- Future Work

Graph Clustering

- Hierarchical
  - *Basuchwdhuri, P., Chen, J., (2010). Detecting communities using social ties, Proceedings of the IEEE International Conference on Granular Computing, San Jose, pp. 55-60.*
  - *Girvan, M., Newman, M. E. J., (2002). Community structure in social and biological networks, Proceedings of the National Academy of Sciences, USA, pp. 8271-8276 .*
- Partitional
  - *Jain, B., Obermayer, K., (2009). Elkan's k-means for graphs, arXiv:0912.4598v1[cs.AI].*
  - *Ferrer, M., Valveny, E., Serratosa, F., Bardaji, I., Bunke, H., (2009). Graph-based k-means clustering: A Comparison of the set median versus the generalized median graph. Lecture Notes in Computer Science, 5702:342-350.*
- Spectral
  - *Liu, J., Wang, C., Danilevsky, M., Han, J., (2013). Large-scale spectral clustering on graphs, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing.*
  - *White, S., Smyth, P., (2005). A spectral clustering approach to finding communities in graphs, In Proceedings of the 5th SIAM International Conference on Data Mining, Philadephia, pp. 76-84.*
- Random Walks Based
  - *Pons, P., Latapy, M., (2005). Computing communities in large networks using random walks, Lecture Notes in Computer Science, 3733:284-293.*
  - *Macropol, K., Can, T., Singh, A., (2009). RRW: repeated random walks on genome-scale protein networks for local cluster discovery, BMC Bionformatics. 10:283.*

Time-evolving Graphs

- ## Clustering Approaches
  - *Sun, J., Faloutsos, S., Papadimitriou, S., Yu, P. S., (2007). GraphScope: Parameter-free mining of large time-evolving graphs, In Proceedings of the ACM SIGKDD International Conference Knowledge Discovery in Databases, San Jose, pp. 687-696.*
  - *Aggarwal, C. C., Yu, P. S., (2005). Online analysis of community evolution in data streams, In Proceedings of SIAM International Conference on Data Mining.*
  - *Semertzidis, K., Pitoura, E., Terzi, E., Tsaparas, P., (2016). Best Friends Forever (BFF): finding lasting dense subgraphs.*
- ## Non-clustering Approaches
  - *Rossi, R. A., Gallagher, B., Neville, J., Henderson, K., (2013). Modeling Dynamic Behavior in Large Evolving Graphs, In Proceeding of the 6th ACM International Conference on Web Search and Data Mining, pp. 667-676.*
  - *Toyoda, M., Kitsuregawa, M. (2005). A System for Visualizing and Analyzing the Evolution of the Web with a Time Series of Graphs, in Proceedings of the 16th ACM Conference on Hypertext and Hypermedia, pp. 151-160.*
  - *Semertzidis, K., Pitoura, E., (2016). Time traveling in graphs using a graph database, in Proceedings of the Workshops of the (EDBT/ICDT).*

Graphs and Labels

- Cluster graphs regarding their labels
  - Cluster comments from online news
    - *Aker, A., Kurtic, E., Balamurali, A. R., Paramita, M., Barker, E., Hepple, M., Gaizauskas, R.., (2016). A graph-based approach to topic clustering for online comments to news, In Proceedings of the 38th European Conference on Information Retrieval, pp. 15-29.*
  - Solution at SRC problem
    - *Scaiella, U., Ferragina, P., Marino A., Ciaramita, M., (2012). Topical clustering of search results, In Proceedings of WSDM-12, pp. 223-232.*
  - Combine labels and time evolution networks
    - *Ferlez, J., Faloutsos, C., Leskovec, J., Mladenic, D., Grobelnik, M., (2008). Monitoring network evolution using MDL, In Proceedings of the IEEE International Conference on Data Engineering, pp. 1328-1330.*

Contents

Communities Detection

- Snapshot: Split the time period into years and construct a a graph snapshot that reflects the state of the network for each year.
- Communities Detection
  - Louvain algorithm
  - Louvain algorithm with multilevel refinement
  - SLM algorithm
- Final goal: Optimize modularity function

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

Louvain Algorithm

- Initial state: Each node of the graph constitutes a cluster
- Consists of two repeated steps
  - Step 1: A "greedy" assignment of nodes to communities, favoring local optimizations of modularity
  - Step 2: The construction of a reduced network by merging the nodes communities found in the first step
  - These two steps are repeated until no further modularity-increasing reassignments of communities are possible

Label Extraction

- Labels are text characteristics of the identified communities
- Labels pre-processing
  - Tokenization
  - Stopwords removal (Natural Language Toolkit list)
  - Stemming (Porter's algorithm)

Stable Communities Identification

- Stability threshold: We define a community as stable between two consecutive snapshots G_i and G_{i+1}, if at least t% of the entities (nodes) that belong in the community in snapshot G_i, also belong in the community at snapshot G_{i+1}.

- Our proposed approach examines the persistence of communities and stable labels by considering specific **time windows**

Stable Communities Identification Algorithm(1)

- Input of the algorithm
  - starting time point
  - ending time point
  - stability threshold
  - snapshots' communities
  - communities' labels
- Output of the algorithm
  - the clusters that persist
  - clusters' labels that continue to appear

Stable Communities Identification Algorithm(2)

- Steps of the algorithm
  - Step 1: Identification of communities with common entities between n and n+1 snapshot
  - Step 2: We examine if the number of common entities is above or equal the given threshold (1 condition)
  - Step 3: We examine if there are common labels that characterize the clusters that fulfill condition1 (condition 2)
  - Step 4: If condition 1 and 2 are fulfilled, the new entities of the n+1 snapshot cluster are added in the pool of the entities of the cluster that continues, and the common labels are kept.

- *The above steps are repeated with input the n+1,n+2, … years and the clusters that continue existing from previous years with their common labels till the ending year is reached*

Contents

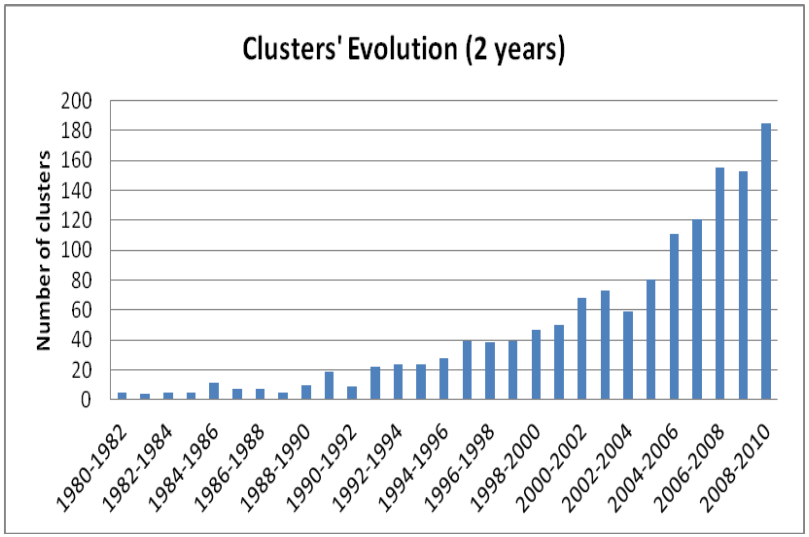
Dataset

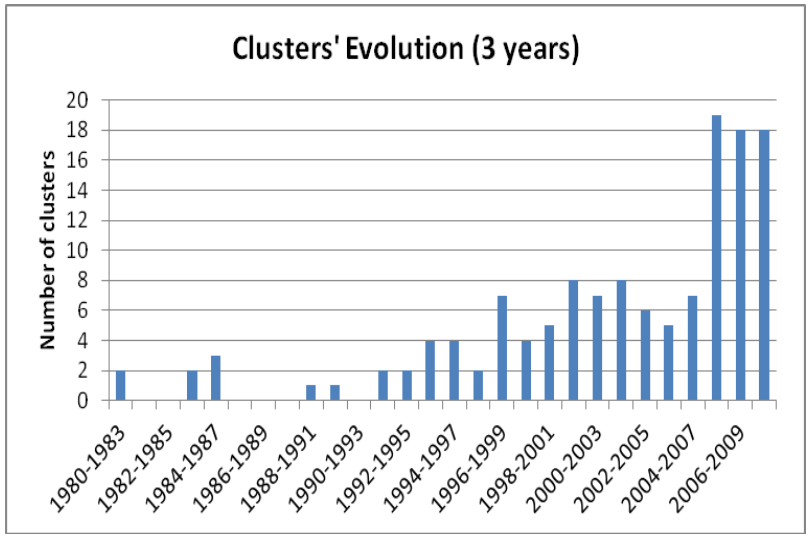
- Dataset
  - Downloaded from dblp site
  - Data from 1980 till 2010
  - Snapshot: Each year from 1980 till 2010
- Network
  - A co-authorship network, is constructed where the nodes of the graph represent the authors, and two nodes are connected if the corresponding authors they represent have a joint publication.
- Labels
  - Publications' titles
- Sliding window resolution ranging from 2 to 6 years
- 1 year stride

Parameters Tuning

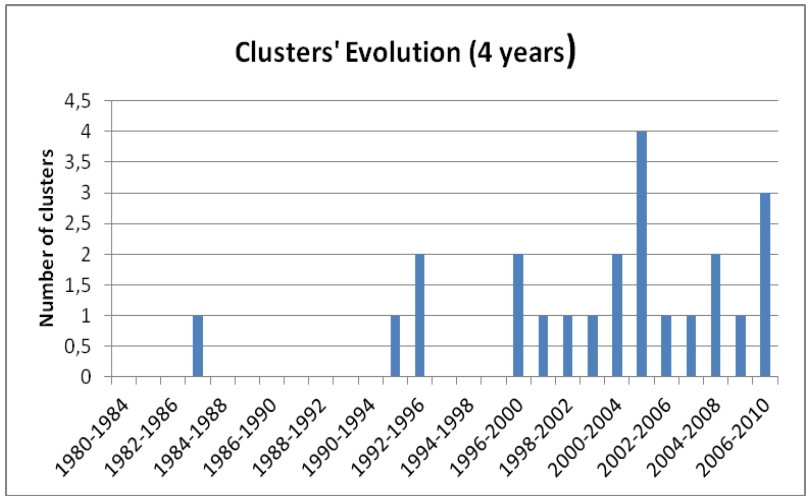
- Clustering algorithm
  - Louvain
  - Louvain with multilevel refinement
  - SLM
- Resolution parameter
  - 1.0
  - 2.0
- Modularity function
  - Standard
    - *Newman, M., Girvan, M. (2004). Finding and evaluating community structure in networks, Physical Review E, 69:026113.*
  - Alternative
    - *Traag, V. A., Van Dooren, P., Nesterov, Y., (2011). Narrow scope for resolution-limit-free community detection, Physical Review E, 84(1):016114.*
- Stability threshold
  - 20%
  - 60%
  - 80%

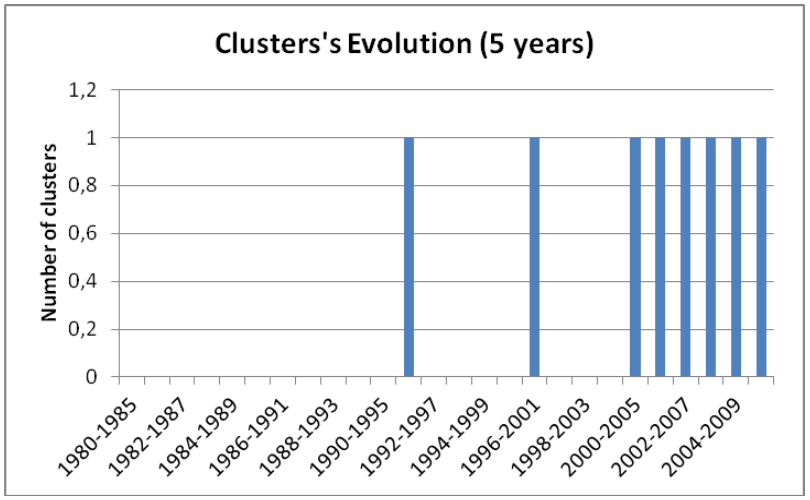
Clusters' Number (1)

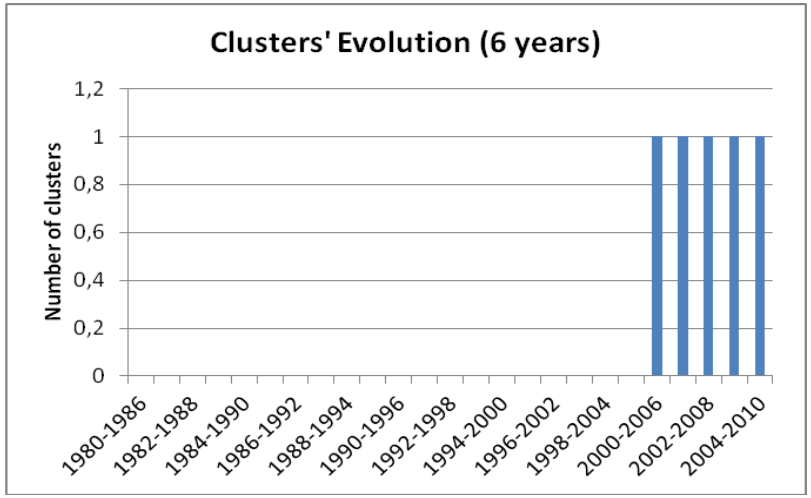




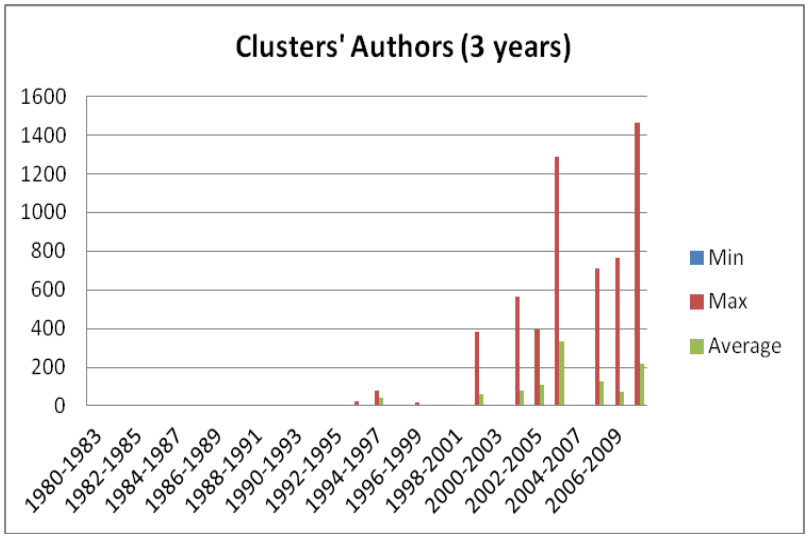
Clusters' Number (2)

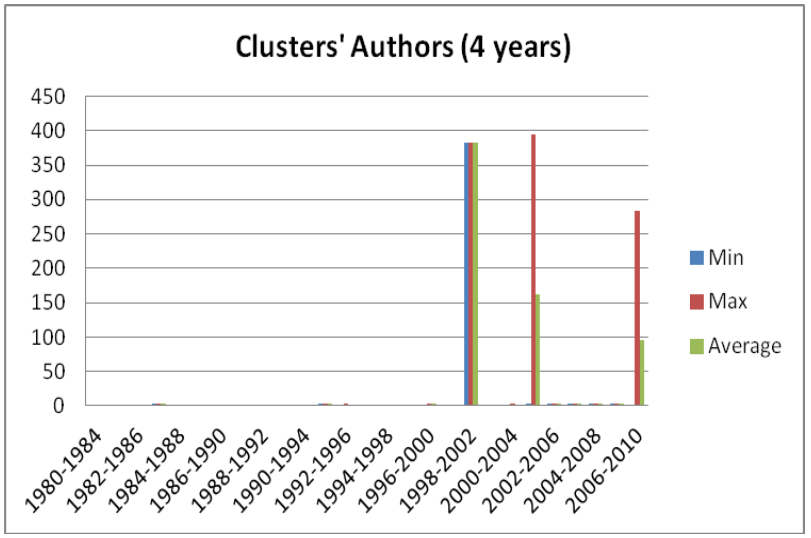






Clusters' Authors





Clusters' Labels (3-year time window)

Time intervals	Labels
1991-1994	(chemistrii, topolog, organ, **graph)**, (reconstruct)
1992-1995	(**graph**, topolog, organ), (**graph**)
1993-1996	(logic), (**graph**), (toppolog, organ, **graph**, theory), (cluster)
1994-1997	(**fuzzi**), (use), (recurs), (reason)
1995-1998	(build, smalltalk), (**graph**)
1998-2001	(methylas, enzym), (**fuzzi**), (noncoher), (distanc, code), (control, fuzzi)
1999-2002	(distanc, code), (**fuzzi**), (inform), (multimedia), (inform), (measur), (induc), (model)
2002-2005	(william, lowel, putnam, mathemat, competit), (network), (use, neural, control, intellig, **fuzzi**), (ultrasound), (fade), (test, data)
2005-2008	(william, lowel, putnam, mathemat, competitt), (**fuzzi**), (network), (linear, **fuzzi**), (invers), (algorithm, cdma), (**graph**), (process), (entropi, hidden, markov, chain, rate), (code), (transmiss), (algorithm), (shrinkag), (stenographi), (orthogon, polynomi), (morphism), simul, n-qubit, quantum, system), (sequenc), (estim)

Top 3 Labels (3-year time window)

Time intervals	Labels	Occurrences at clusters	Occurrences at time intervals	Occurrences at 1980-2010
1992-1995	graph	4, 6	2128	19747
	topolog	12	259	3434
	organ	7	187	2557
1993-1996	logic	6	1466	10668
	graph	4,5	2358	19747
	theori	5	1008	9873
1994-1997	fuzzi	7	1287	14733
	recurs	6	365	2671
	reason	4	530	3478
1995-1998	build	6	308	2852
	smalltalk	9	27	67
	graph	4	2568	19747
1998-2001	fuzzi	7, 7	2970	14733
	code	4	6760	17328
	control	6	4685	30250
1999-2000	inform	13	4140	24646
	fuzzi	22	3058	14733
	model	6	8775	59363

Clusters' Labels (4-year time window)

Time intervals	Labels
1991-1995	(topolog, organ, graph)
1992-1996	(topolog, organ, graph), (graph)
1996-2000	(methylas, enzym), (block)
1997-2001	(methylas, enzym)
2000-2004	(william, lowel, putnam, mathemat, competit), (match)
2001-2005	(william, lowel, putnam, mathemat, competit), (intellig, control, use), (test), (network)
2004-2008	(william, lowel, putnam, mathemat, competit), (stenographi)
2006-2010	(william, lowel, putnam, mathemat, competit), (messag), (relai)

Top 3 Labels (4-year time window)

Time intervals	Labels	Occurences at clusters	Occurences at time interval	Occureneces at 1980-2010
1991-1995	topolog	14	308	3434
	organ	8	224	2557
	graph	7	2504	19747
1992-1996	topolog	14	352	3434
	organ	8	242	2557
	graph	5, 7	2816	19747
2000-2004	mathemat	5	589	2786
	competit	5	565	2049
	match	6	978	4832
2001-2005	test	18	2385	9315
	control	9	7882	30250
	network	22	11323	53321
2004-2008	stenographi	6	80	199
	mathemat	5	936	2786
	competit	5	710	2049
2006-2010	relai	10	1319	1458
	matthemat	5	1196	2786
	competit	5	884	2049

Contents

Conclusion

- There are communities that persist over time with stable characteristics
- As the time interval increases the number of communities that persist over time decreases rapidly
- In the biggest time interval, 6 years, only a few clusters managed to persist with at least one common label throughout all these years
- After 2000 the research community is more active and as a result more clusters persisted and met the prerequisites of the proposed methodology.
- In the 3-year time window it was found that there are labels that keep existing over time but at different clusters.
- The occurrences of the top 3 labels that appear at 3- and 4 year time intervals in the whole time window showed that actually there exist keywords that keep being of interest to the overall research community over time but they do not appear continuously in the same cluster.

Contents

Future Work

- Improve algorithm's efficiency and scalability by using appropriate indexing structures or the design of pruning methods
- A more refined approach regarding community labels can be adopted
  - Group together the labels that concern the same topic maybe with a clustering approach
  - Ontologies
  - Extract and exploit as labels the keywords from the abstract or from the main body of the paper
- The stability threshold could be adaptive at the size of the cluster.
- Repeat the experiments after a few years with a more recent time window
- Test different clustering algorithms
- Relax the clustering approach by tracking communities by using a measure of their similarity or relevance.

Thank you!