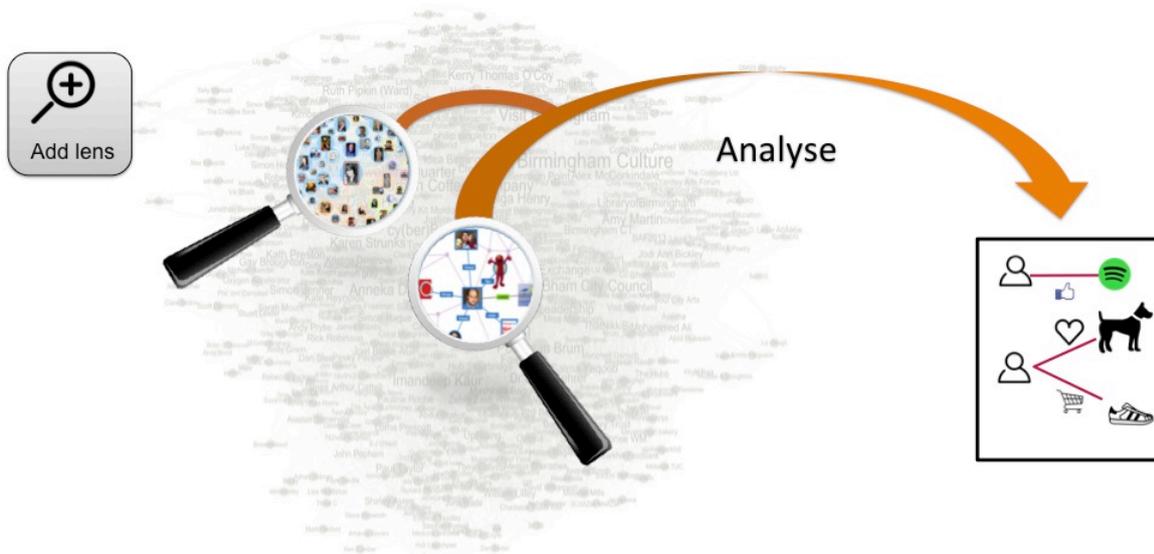# What is hidden in my network?
Interactive Exploration of Attributed Graphs

Project description

Graphs are widely used for the representation of social networks, gene and protein interactions, communication networks, or product co-purchase in web stores. In this project we study **attributed graphs** in which nodes and edges have additional attribute information (e.g. age in a social network). Attributed graphs are very common as they model both relationships (edges) between objects (nodes) as well as object and relationship descriptions (node/edge attributes). However, analyzing such complex graphs with millions of nodes/edges and thousands of attributes is very hard. Typically, the graph is storing all types of information, while one specific user is only interested in a subset of this information. Inferring such subset is still an open challenge.

We aim at an interactive exploration of attributed graphs, assisting the user by automated methods. We provide portions of the graph that seem relevant to the user interest. In this project, we aim to learn the relevant subgraph by modeling the user interests. A viable way to analyze your own network and retrieve important summaries would be to provide *example nodes*. These examples act as magnifier lenses on the the portion of the graph you perceive as relevant.

For example, as a user of a social network you want to analyze your own network and find users that share similar interest with you. You only know a small set of interesting users, so you provide them to our system as examples. The system learns the similarity of these users, analyzes the residual network, and returns the set of common structures found (e.g. they are all subscribed to a music channel). Also e-commerce companies might be interested in such technology. Suppose that Amazon wants to sell new interesting products and wants to know what are the habits of the users they target. They know all the users that bought a specific product and they know the connections between users. Using example Amazon looks at a smaller, but significant portion of the graph and finds that, for instance, north American users buy large shoes and love dogs.

In this project we will study how to learn a relevant subgraph in attributed graphs given an input set of examples (nodes or structures) as initial seeds. There is quite a few numbers of solutions for similar problems, such as finding similar structures [1], or performing ad-hoc graph clustering using nodes as seeds [2]. However, the problem we aim to solve is more general in that we want to infer a relevance score from a set of examples supplied by the user. The relevance score will allow the user to effectively restrict the search space and either visualize the results or perform further analyses on the data.

The challenging part will consist in the definition of a measure for the relevant subgraph to achieve both efficient computation *for large graphs* and accurate results *for complex patterns*. The approach will be experimentally evaluated on several real graphs at our disposal, e.g. Amazon co-purchase graph, Freebase knowledge graph.

Project goals

1. Read and understand state-of-the-art in subgraph mining literature.
2. Formalize a quality measure to retrieve the relevant portion of the graph.
3. Develop efficient and accurate algorithmic solutions for this interactive setting.
4. Evaluate the proposed methods using large graphs at our disposal (e.g., Amazon, Freebase) or online available (e.g., `http://snap.stanford.edu/`, `http://konect.uni-koblenz.de/`).
5. Write a submission-ready scientific publication for a top data mining conference/workshop.
6. If successful, present the work at an international conference.

Required skills

- Good programming skills (preferably Java and C++)
- Basic data mining concepts, are beneficial but not required.
- Interest in graph mining concepts and innovation of new algorithmic solutions.

Team and dates

The team will consist of 3-5 students. The project will start mid April, 2016. A preliminary meeting (February 8th, 2016 at 2-3pm, Office E1-02.2) will be held to answer potential questions.

Contacts

Please contact us to apply for the project or ask for more information:
- Prof. Dr. Emmanuel Müller (`emmanuel.mueller@hpi.de`) - Office: E1-01.2
- Dr. Davide Mottin (`davide.mottin@hpi.de`) - Office: E1-02.2

Bibliography

[1] D. Mottin, M. Lissandrini, Y. Velegrakis, and T. Palpanas. Exemplar queries: Give me an example of what you need. *PVLDB*, 7(5), 2014.

[2] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller. Focused clustering and outlier detection in large attributed graphs. In *SIGKDD*. ACM, 2014.