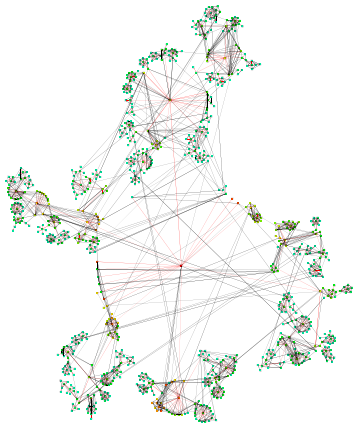


Figure 1: An Amazon recommendation graph

Data sets

Amazon co-purchasing Networks

The webshop Amazon [1] offers a SOAP interface by which data that is available on their webpages is also freely available for automatic requests. In theory, every information on the websites should be available by this interface but it turned out that the information is not necessarily the same. **Andreas Gerasch** thus built a web bot that crawls the webpages as if it were a normal user. The bot starts at some predefined starting book, indicated by the unique ISBN number. As additional parameters, the number of threads and the depth unto which links are followed are given.

The bot searches for links to other books that are placed directly under the text line: 'Customers who bought this book also bought', the co-purchasing information. There are no more than six links under this title (August 2006), and all of them are stored in a data base. This data base stores each search with a different ID such that it is possible to store the same book multiple times. The tool built by Andreas provides many helpful features such as building the difference graph given two graphs G_1 and G_2: In a difference graph, every vertex or edge that is only in G_1 is colored red, every vertex or edge that is only in G_2 is colored green, all other vertices and edges are black.

Figure 2: A very small part of the Netflix data.

Netflix

Netflix is an internet base video rental company. The users can evaluate the films they have already seen and based on this information, a recommender system called CineMatch makes recommendations to the user of what to see next. Therefore, CineMatch tries to calculate how a given user will rate a given film and its RMSE value (root mean square error) is around 0.957. Netflix has published a part of its data set, consisting of around 100 million evaluation events between 500,000 users and 17,770 films and set out a prize of $1,000,000 if anyone can improve CineMatch by 10 percent points. Find more information under: www.netflix-prize.com.

Autonomous System

The National Laboratory for Applied Network Research (**NLANR**) has documented the evolution of the Internet from November 1997 to March 2001 and made this data publicly available at *http://moat.nlanr.net/AS/*. An *Autonomous System* is generally a group of routers and computer networks under the control of one entity.

The raw data provided by the NLANR gives routing information that implicitly contains information on which Autonomous Systems are directly connected. For each month, the data of the first ten days is combined and displayed as a network, where the Autonomous Systems are represented by the vertices and two vertices are connected if there is at least one path in which the two Autonomous Systems are listed consecutively[1]. Note that this data is not perfect: If some paths are only rarely used, an edge might emerge in one month, be

[1]Our thanks go to Jan Vitense who provided us with this data

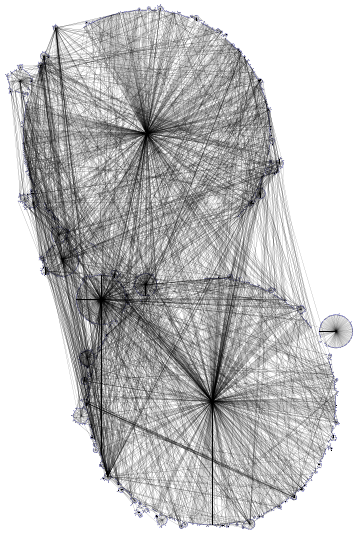


Figure 3: The AS network from November 1997.

missing in the next few months, and show up again later. We thus regarded only those edges and vertices as *new* that had never before been seen in any of the earlier networks but where both vertices had already been in at least one of the preceeding networks.

Co-Authorship Network

The data was kindly provided by M.E.J. Newman who used the same networks [2, 3]. These networks are available from his homepage http://www-personal.umich.edu/ mejn/netdata.

The networks represent authors of papers published in the online preprint archive *arxiv*. In the first data set, all papers published between 1995 and 1999 were analyzed, and in the second, all papers published between 1995 and 2003. The networks are weighted but in this analysis we disregarded the weights, and only the biggest connected component is used. The first data set contains 13,861 vertices and 44,619 edges, and the second 27,519 vertices and 116,181 edges. Of those, 57,277 edges are new edges between authors that were already present in the first network. These edges are used to determine the new edge distance distribution.

Word Association and Protein-Protein Interaction Network (PPI)

The data was kindly provided by Palla et al. who used the same networks in [4]; these networks are available together with the tool *CFinder* at http://angel.elte.hu/ vicsek.

For the word association network, people were asked what word they associate with a given word, and two words are connected by an edge if at least one person associated the two words with each other. The network contains 7,205 vertices and 31,783 edges in the biggest connected component which was used here. Palla et al. name a website from which the data can be obtained that seems to be outdated. The data can now be found starting from page http://w3.usf.edu/FreeAssociation/intro.html.

The protein-protein interaction network presents proteins from the organism *Saccharomyces cerevisiae* that were found to interact with each other in biological experiments. This network contains 2,445 vertices and 6265 edges in the biggest connected component, where self-loops have been removed.

Open Problems

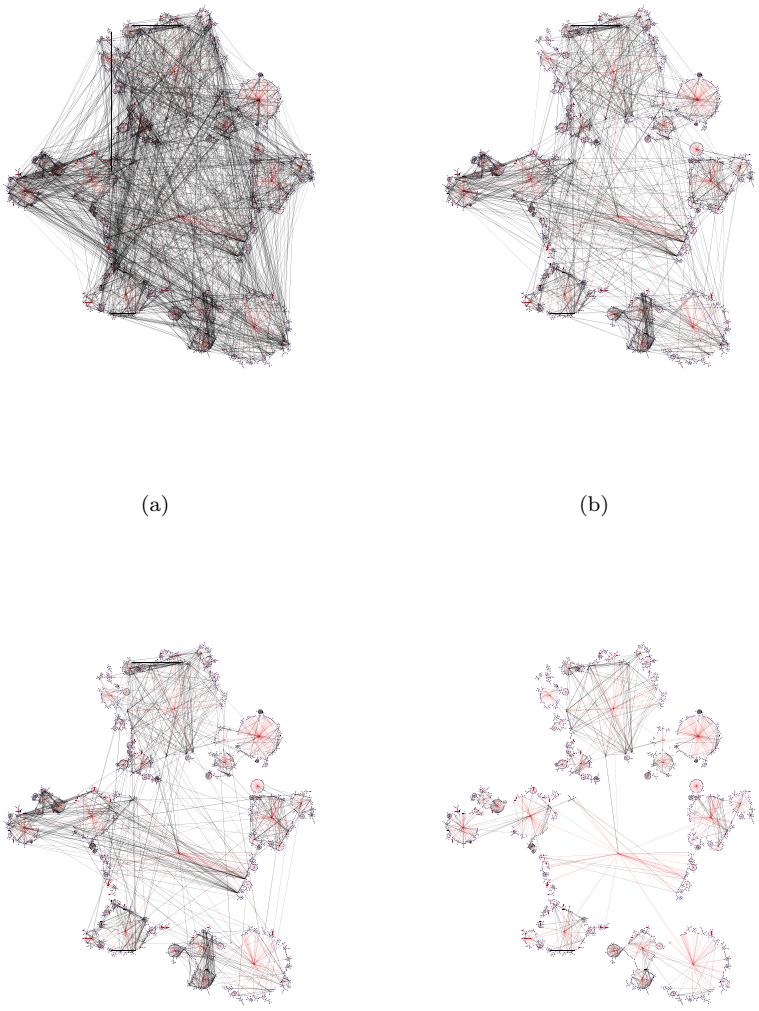
1. Let G be a large graph in which most of the edges are within dense subgraphs. Let T denote a spanning tree and for every edge $e = (v, w) \notin T$ define its *tree distance* $tdd_T(e)$ as the distance of v and w in T. Find the spanning tree with minimal tree distance sum, i.e.,

$$\min_T \{\sum_{e \notin T} tdd_T(e)\}. \tag{1}$$

   This question is equivalent to finding the *minimal length strictly fundamental cycle base* of a graph. Open question: is the minimal length in $O(m \log^2 n)$?

2. Let G_k denote a metagraph that represents all k-cliques of a given graph G. Connect two k-cliques if they share $k-1$ vertices and compute all connected components. Each component C defines a *community* as the set of all vertices contained in at least one k-clique in C. Note that vertices can be contained in more than one community. How (when) can the graph be drawn such that all communities are in a contiguous area without including non-community vertices? This question should be related to the question of when Venn-diagrams can be drawn. Can unit-disk-graphs or other geometric graphs always be drawn like this? Can the number of non-contiguous areas be minimized? Approximated?

3. What's the best method to cluster a bipartite graph?



(a) (b)

(c) (d)

Figure 4: The PPI network of yeast. (a) All 6448 edges; (b) 5234 edges (81%); (c) 4909 edges (76%); (d) 3877 edges (60%).

References

[1] www.amazon.com, www.amazon.de.

[2] Mark E.J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, USA*, 98(2):404–409, 2001.

[3] Mark E.J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69:066133, 2004.

[4] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814, 2005.