

Dynamic Social Networks and the Textrend/CIShell Framework

George Kampis^{1,2,3}, László Gulyás^{1,2}, Zoltán Szászi^{2,3}, Zalán Szakolczi^{2,3}, and Sándor Soós³

¹Collegium Budapest, Institute for Advanced Study (www.colbud.hu)
H-1014 Budapest, Szentháromság u. 2.

²Department of History and Philosophy of Science, Eötvös University (<http://hps.elte.hu>)
H-1117 Budapest, Pázmány Péter s. 1/C

³Universitas Press (www.unipresszo.hu)
H-1055 Budapest, Markó u. 7.

Corresponding author: George Kampis gkampis@colbud.hu

Abstract:

Most of the work in Social Network Analysis has been devoted to static networks. However, real-life networks are essentially dynamic, and there is an increasing interest in their inherent features: in how time-dependent equivalents of classic network measures can be defined for descriptive analysis, or how dynamics relates to time constrained samples in a cumulative network. We present results from a research project in temporally aware text-mining, aiming to develop novel, dynamic network measures and a toolkit that implements them. The measures offered range from simple time series of classic global network statistics, to methods for observing the change in node-level measures (e.g., centrality) and to automatically following the development of the “most important” nodes, or to the capturing of cluster morphosis and evolution. The TexTrend framework is built upon the CIShell cyberinfrastructure shell and provides a platform for the dynamic analysis of both structured and unstructured textual information.

1. Introduction: Dynamic Network Analysis

The development of the theory of social and indeed all kinds of networks has been in the foreground of interest in the past decade. Many network characteristics were identified and several network models were introduced. However, the overwhelming majority of these have dealt with networks understood as static objects.

For example, the various well-known measures such as diameter, path length, or the various forms of centrality are defined on invariant, and by that virtue, static networks, and also the study of the most important statistical properties such as degree distributions assume the same form. Real world networks are, however, inherently dynamic, as virtually no natural network pops up instantaneously. A few issues of network generation have been dealt with by network growth models (e.g. Barabasi and Albert 1999, Pennock *et al.* 2002, Barabasi 2009), but the aim of these was to characterize properties of the resulting network (i.e. the envelope), and not the dynamic process itself. Also, in real networks the changes often occur for endogenous, network-internal reasons that are more complicated than can be grasped by a simple growth rule: for instance, the change of a node's internal attributes can imply changes in its connectivity, and so on. (In an epidemic network, the recovered patient does not infect any more, and no more infection links are formed from this node.) Interests of this kind put questions of the interrelated dynamics *of* and *on* networks (e.g. Kossinets and Watts 2006, Shalizi *et al.* 2007, Leskovec 2008, Ganguly *et al.* 2008) into a new focus of investigation. Several important problems ranging from virology (Barrat *et al.* 2008), to contact networks (Sloot *et al.* 2008, Sloot *et al.* 2009) and ecological theory (Jordan and Scheuring 2004) similarly invite a treatment of dynamic networks on their own. A list of early works is given in (Carley 2004).

Partly as a response to these current challenges, several approaches and tools were introduced recently (Trier and Bobrik 2007, Falkowski *et al.* 2007, Lahiri and Berger-Wolf 2008, etc.). These are mostly isolated works, however, focusing on special problems. Only a few aim at developing general methodology. The SONIA project of Stanford (Bender-deMoll and McFarland 2006) promises the visualization of longitudinal network data but also raises some of the most important methodological questions for general dynamic networks. New problems include that of slicing, i.e. the use of thin or thick time slices or windows, and the implied question of how a flexible binning (the grouping of raw data into varying time slices) is achievable to obtain relevant network properties – at a zero width snapshot the network properties, at a full time window, the dynamics is lost (Moody *et al.* 2005). SONIA is built on the assumption that time stamped and -sliced data will grow in significance. Yet public longitudinal network data are still rare (a comprehensive list is Sonia 2009).

Dynamic Network Analysis (DNA, Carley 2003) is the name of a prolific approach to various Social Network Analysis problems (the general name is slightly misleading as DNA is more about modeling and decision support; also, DNA considers special nodes which, just like human agents, can learn). DNA uses the meta-matrix concept (Carley 2003) to grasp multi-mode, multiplex, dynamic networks with multiple node attributes (such as a person's relation to events, organizations, resources, and other people). Some radically new network measures such as the notion of “cognitive load” were introduced – this is the amount of

information that an agent has to handle in order to maintain its position in the network, and translates to connection intensity in a multi-net. DNA also endeavored to develop an interoperable dynamic network toolkit (Carley *et al.* 2005), the arrival of which is yet to be seen. All in all, this is another interesting endeavor based on the recognition of the entirely new kinds of problems encountered in the dynamic domain, but it does not alone change the status of the latter.

The above (listed and not listed) works signify important first steps in what promises be a long development. Despite the recognized needs, currently there is no available set of methods or even software toolkit that would be comparable, in its grasp or generality, but also in its grounding in theory, to the ones at hand for classical network analysis. The latter kinds are too numerous to be listed here; Wikipedia (2009) lists some 40 major network analysis tools and a few further repositories.

The aim of our work is contribute to change in the situation of dynamic networks. Our motivation comes from two problem domains. One is text mining and trend analysis (www.textrend.org) where we focus on temporal changes in textual corpuses such as blogs or scientific publications, and the other, more recent, is dynamic network modeling and the analysis of multi-scale, multilevel natural systems (www.dynanets.org), where we deal with temporal changes in networks as consequences of various interacting levels of node dynamics, such as in epidemic networks of HIV infections. A framework of pre-existing network tools, of closer interest for our current work, is the Network Workbench (NWB, <http://nwb.slis.indiana.edu/>), which is the first truly interoperable and open, generally accessible and extendable, integrative software toolkit for network problems. We use NWB, or more precisely its core shell, to accommodate our new and self-integrating software modules that realize generic dynamic network tools.

In this paper we introduce a few dynamic measures and their realizations in a compact software module (a script library), we discuss the place of the latter in the TextTrend trend analysis toolkit that uses NWB's Cyberinfrastructure Shell, and finally, we show some details of a sample application in Social Network Analysis viz. bibliometrics.

2. A Dynamic Network Analysis Toolkit

2.1. The Ambition

In the long run, we envision dynamic network analysis methods that are as widespread and highly developed as their static counterparts. This defines the global context for the investigation. Our more local aim here was to develop a set of new dynamic network measures and to implement them as a programming tool, to allow the user to start analyzing networks dynamically; these are our first steps in the chosen direction.

2.2. New Measures

We developed several new concepts and implemented them in a package, called Dynnet, which uses the **R** statistical computing environment (<http://www.r-project.org/>); the latter is, in turn, integrated into the TexTrend system. We use three different approaches.

(1) The global level approach. At the global network level, we focus on the changes of the whole network, such as the alteration of the diameter of the graph. We perform very simple calculations here. We calculate the static measures of the network at every moment, and then visualize these values in the function of time.

(2) The local level approach. At the local level, we examine all or some selected local nodes of the network. There can be many local statistics (such as local clustering or assortativity); of these, at the moment Dynnet only focuses on centrality. The idea is that a graph may have a high number of nodes, yet we want to concentrate on just some important ones. For the determination of node importance we use the classic centrality measures and we analyze the dynamics of the selected nodes.

We also added stability measures to express how stable the environment of these selected nodes is. To that end, we examine the dynamics of the neighborhood of a central node: the permanent neighbors of the node, the changes in the strength of its connectivity, the average link strength, and the monotony trends in the link fluctuations. Essentially, we consider a node as stable if it does not change neighbors and also its link strengths do not fluctuate too much (clearly, if you keep your neighbors but the link strengths drop, it is like losing them).

(3) The cluster level approach. Finally, we want to study the evolution of network clusters. We want to find out if a cluster has developed from another one in a previous time slice. We use different clustering methods to separate communities in the network and we introduced various computational functions to find the “same” clusters during dynamic evolution. (Visualization of the evolution of communities is still under development.)

2.3. The Dynnet Package

The package consists of a set of R scripts, implementing the functions of the above mentioned three approaches (Szakolczi 2009, Dynnet 2009). The package was prepared to visualize dynamic network trends; accordingly, most Dynnet functions can plot diagrams.

First we build (or read in) a set of networks as an iGraph list (that is, a list of networks) using the iGraph network analysis library in **R** (<http://igraph.sourceforge.net/>). The Dynnet analyzing functions expect this list, together with a list of years (or months, or any other trend parameters; for simplicity, in the following we assume that the trend parameter is indeed time, measured in years). Below we discuss the operations of the main functions, and demonstrate the results on a few examples.

2.3.1. Global level analysis.

These methods are indeed very simple. We calculate the classical statistics of the network in every static time slice and visualize them together in the function of the trend parameter. The statistics we implemented are:

- Node count
- Edge count
- Degree distribution
- Diameter
- Average path length
- Number of complete subgraphs
- Group cohesion
- Density

Calculations of these are straightforward function calls in **R**; Dynnet only executes and assembles their series into composite figures. As an illustrative example, consider Figure 1. and Figure 2. which present the diameter and average path length plots for the time slices of a selected conceptual network in the period 1993 – 2008 (taken from a Web of Science corpus of papers and keywords with the word “intentionality” in the title).

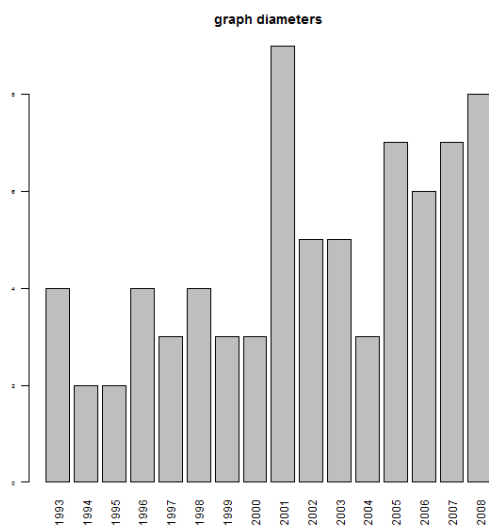


Figure 1. Dynamic diameter of the networks

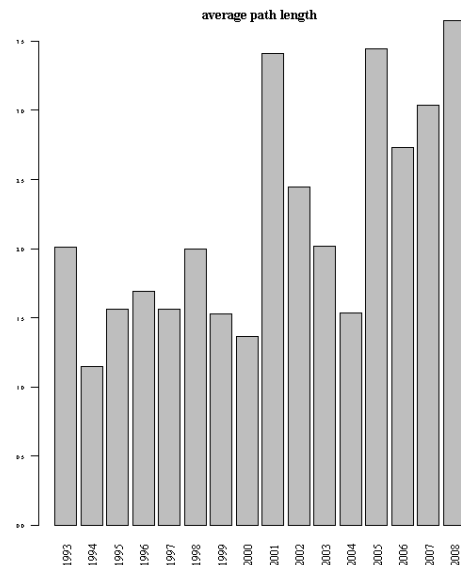


Figure 2. Average path lengths

2.3.2. Local level analysis.

As mentioned above, in this local level analysis, we study some selected nodes, namely those which are characterized with high centrality scores. The centrality score is a specific

value to estimate the relative node importance within the network. Many centrality measures exist, and they are calculated differently. The centrality measures we apply are the following:

- Betweenness centrality
- Closeness centrality
- Eigenvector centrality
- Kleinberg centrality
- Page rank scores (using the Google page rank algorithm)

We again use the standard iGraph functions of **R** to calculate these centrality scores for every static time slice. The Dynnet module first obtains these results and uses them as starting points. After this initial calculation, we select the largest centrality scores according to a chosen threshold. From then on, we concentrate on the remaining nodes' dynamics. We identify which ones show high centrality scores in more than one year; this way we can identify the nodes which are important in a more permanent fashion.

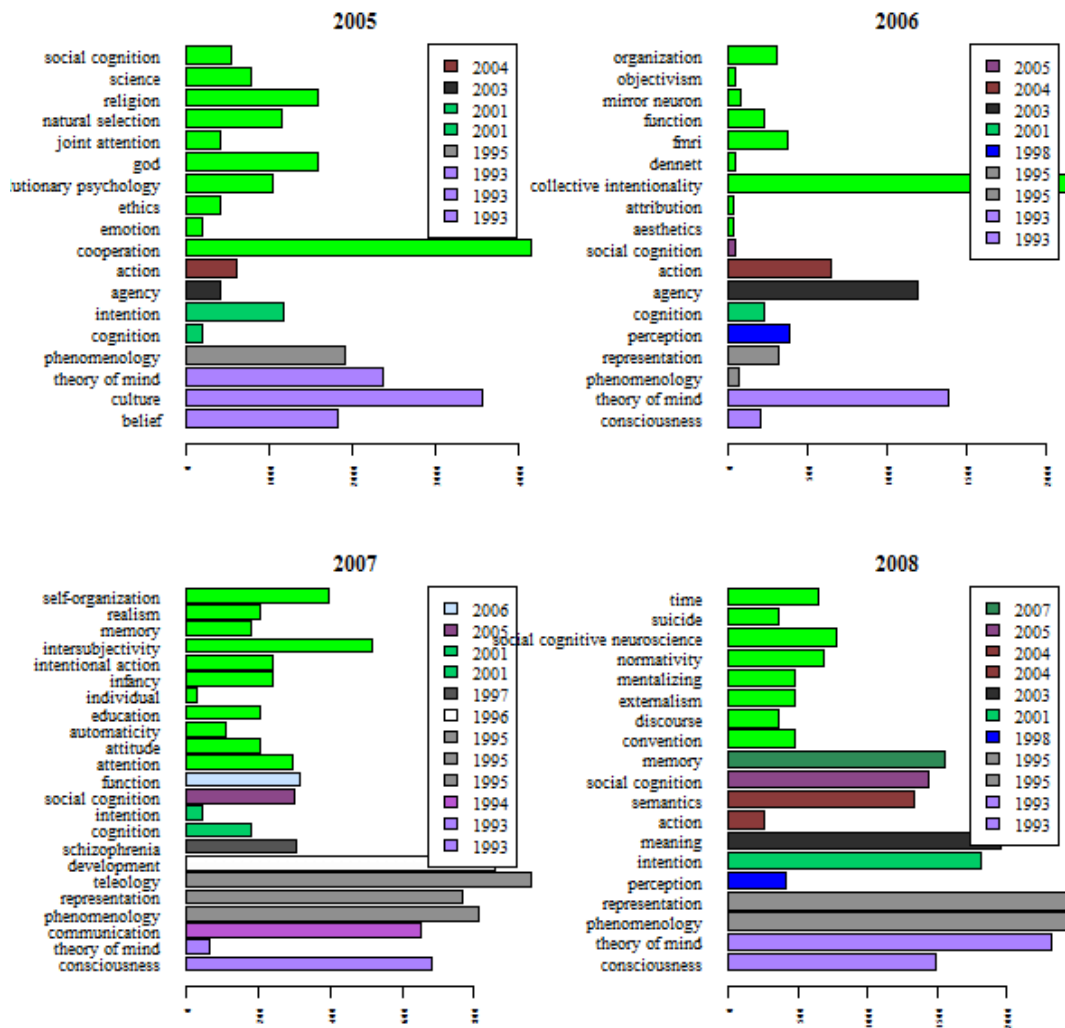


Figure 3. The dynamics of betweenness scores of the leading nodes

Figure 3. shows an example (from the same corpus as above) with nodes having higher betweenness centrality scores in the dynamic sense. Results are color-coded. Green means that a given node is “new” in the plot, that is, the given node has a high centrality value for the first time in the given year (i.e. newly important node). Other colors refer to a specific year each; namely the one in which a high centrality value is detected for the given node for the first time.

2.3.3. Stability measures and the dynamic environments of central nodes.

We also want to analyze the dynamics of the neighbors of a given central node. The idea is to first identify the nodes which are the most important in the whole studied interval of the network dynamics (note that so far we have only considered series of snapshots.) Then we relate these all-important nodes to each other in various ways.

We assign a value to every node to express the dynamic weight of the centrality scores of the node. The value is calculated as the sum of the centrality scores of the node in the different years, weighted by the age of the node. (The age of the node is the number of consecutive years in which it has above-threshold centrality values.) Then, we select a “winner node” – the node with the largest dynamic centrality weight. Now we discard nodes with weights below a proportional threshold. The remaining nodes are the ones with the highest centrality values and which are relatively permanent. Hereafter we will call these nodes “the central dynamic nodes” (CDN).

We examine the neighbors of the CDN nodes. If the CDNs have links to each other in more than one year, we can plot a CDN graph. This CDN graph can be said to express a sort of dynamic skeleton of the network. An example CDN graph is presented in Figure 4.

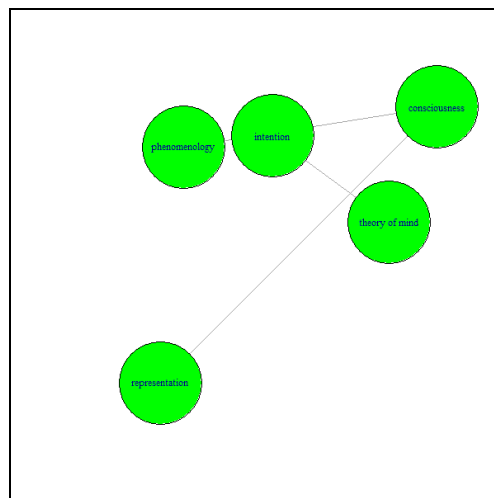


Figure 4. A central dynamic node (CDN) graph

CDNs open room for further analysis. In a next step select those nodes, which were connected to a CDN for more than one year (or more than two years, etc. depending on the selected

threshold). These nodes signify the relatively permanent neighbors of an individual CDN. We compute and visualize these neighbors for each CDN, and in these diagrams we also express the dynamic trends of the link strengths. In this way we can calculate 3 diagrams altogether, for every CDN. Figure 5. shows an example, explained below.

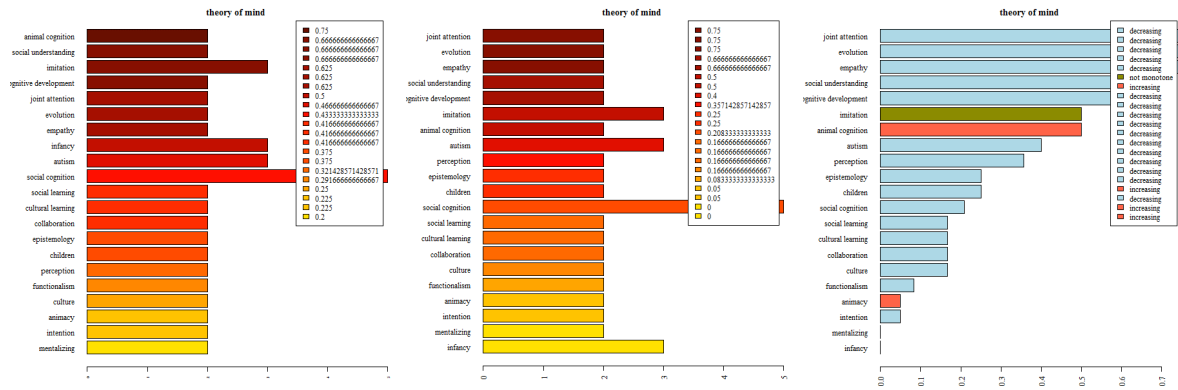


Figure 5. Dynamics of the neighbors of a CDN

In the first diagram, the height of each (rotated) column shows the number of years in which a listed neighbor stays connected to the given CDN. The color of the column refers to the average strength of connectivity (AVC) over these years. In the diagram the neighbors are shown ordered by the AVC values. Colors go through the spectrum from yellow to the red, the latter meaning higher values.

In the second diagram, the heights of the columns again show the number of years in which a given neighbor has a connection to the selected CDN node. Here, however, color represents the average magnitude of fluctuations (AVF) of the link strength between consecutive time slices. AVF values are also ordered in a decreasing fashion. (We use the same spectrum to express AVF as AVC.)

Finally, in the third diagram, the heights of the columns are AVF values, and colors express the monotonicity of the yearly fluctuations of link strengths. If the strength of the links tends to become smaller, the color of the given column is blue. If the connectivity strength is growing over the years, then the color of the column gets the color tomato (meaning increasing link strength). Finally, if the strength of the links fluctuate during the selected years, the column is colored green. On Figure 5, we see a CDN which is losing significance: most of its links decay over the examined period

Our next aim was to calculate stability measures for the CDN-s. We consider a CDN highly stable if its neighbors stay approximately the same over the years and the sum of the AVF values between the neighbors are not too big either (meaning that the variation in the neighbors' connection is not very high). We use three different measures, as different demands may require.

- **M1** Under this (time global) measure, a high stability value means that neighbors are unchanged or recur again, not necessarily in consecutive time slices, but between any two moments. This measure is calculated as the sum of the frequency of occurrence of those CDN neighbors which appear in more than one year in the environment of the CDN, divided by the average value of fluctuation (AVF) of the link strengths between these nodes.
- **M2** In this (time local) measure we consider a central node stable, if its neighbors are identical in two consecutive years. We define this measure as the Jaccard-distance of the set of neighbors in the two consecutive years.
- **M3** On this measure we divide **M2** by the sum of the fluctuation values of link strengths in each year.

The Dynnet package can calculate and plot these stability measures for each CDN. An example is shown in Figure 6. The rows are nodes of the CDN, the (rotated) column heights show **M1**, **M2** and **M3** for the respective nodes in the subfigures from left to right.

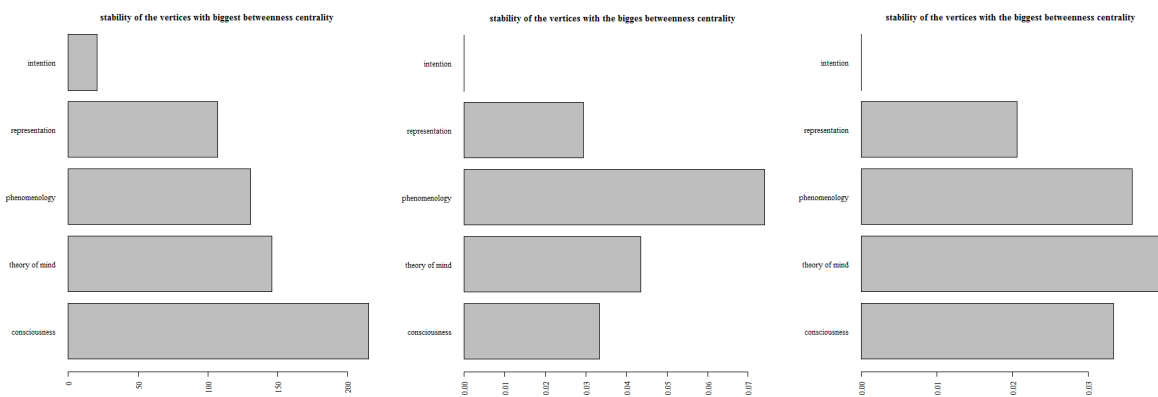


Figure 6. Stability values for a CDN.

In what we discussed so far, we have used betweenness centrality; Dynnet can do the same local level analysis using each aforementioned centrality measure.

2.4. Cluster level analysis

Finally, it will be very useful to separate network communities and to analyze them dynamically. (This is work-in-progress.) We use different clustering methods to start with:

- Walktrap clustering
- Eigenvector clustering
- Betweenness clustering
- Fastgreedy clustering
- Label propagation clustering

As before, again we use the built-in iGraph functions of **R** to find these communities in the individual static networks, at every time slice. Our aim is now to express the time evolution of the resulting communities.

A goal is to plot an evolution tree for the network clusters. To this end, we mark every cluster with an identifier node which is the one with the highest betweenness centrality. We say that a community evolves from another if the new community has the same identifier node as the old. We can use color-coding for lineages: if a community is evolved from another in one step, then the color of these communities will be the same and there will be an edge from the older community to the newer.

We also want to express some statistics of the communities; such as the number of nodes within the cluster, the ages of the cluster, etc. Also for visualization, if there are two communities, and the one has 10, the other 100 nodes, then the size of the first community plotted in the result should be smaller than the size of the second, etc.

We successfully devised the main function which divides the clusters of a given time slice into different parts according to which cluster evolved from which one in an earlier slice. However, the visualization of the cluster evolution as well as the statistical analysis is still under development (hence no figure is shown here).

3. The TextTrend framework

3.1. The Aim

The main aim of the TextTrend framework is to provide integrative, easy-to-use tools for advanced composite text mining/SNA tasks which are characterized by complex analysis pathways, typically not accessible to non-specialists such as data or problem owners. We focus on dynamic information in a double sense: to develop the toolkit using time- (trend-) stamped representations (that can be processed using traditional tools), and to embed novel, innovative dynamic analysis tools in it, to make new kinds of analyses possible.

3.2. CiShell, OSGi, integrative platforms

The TextTrend framework is built on the OSGi (<http://www.osgi.org/>) based CyberInfrastructure Shell (<http://cishell.org/>) platform, which was developed by Börner's group at Indiana University (Börner *et al.* 2003, Börner *et al.* 2007). OSGi is a Java (<http://java.sun.com/>) based service approach, which defines a well structured interface for individual services (modules) that communicate with each other. In this framework, CiShell provides a highly useful additional, integrative layer that defines algorithms as its building blocks, manifested as plugins. In CiShell an algorithm is a functional unit, such as a stand-alone executable, a Java library, or a Java program. An algorithm takes input data, produces output data and can receive any type of parameters. In the context of CiShell, OSGi services are apprehensible as containers and managers of algorithms, so the of managing input/output

data and parameters of algorithms is implemented as a service, and additionally the conversion and validation of data is possible as well.

TextTrend extends the CiShell framework with several individual plugins and also some functionalities such as a workflow service (under development). The workflow service can interpret a previously constructed XML file, which describes a fixed workflow (i.e. pipe) of algorithms. It is important to note that the Java environment is platform-independent, so the OSGi/CiShell/TextTrend solution can offer support for several platforms such as Linux, Windows and Mac OS X. (For TextTrend, Linux and Windows are the primary targets at the moment, but Mac OS X support is planned). All of the used technologies are open-source and free to use, therefore, our developing model is free and open as well.

CiShell (and hence TextTrend) features an interactive, menu-driven, easy-to-use graphical user interface to make analysis intuitive and accessible to the non-specialist. The TextTrend system and its plugins including Dynnet are available at www.texttrend.org.

3.3. The TextTrend architecture

The core of the Texttrend architecture is a set of newly developed CiShell plugins that implement the various steps of a text mining and network analysis process, with a focus on the dynamic features. TextTrend plugins belong to two basic types: library and application plugins. The library plugins integrate powerful frameworks (of free licence), such as UIMA, WEKA, and the **R** statistical analysis project. Application plugins implement functional modules using the former.

UIMA (Unstructured Information Management Architecture) is a general architecture for building and using NLP (natural language processing) functions called Analysis Engines (AE-s). Our application plugins implement various AE-s using the TextTrend UIMA library plugin. Likewise, WEKA is a powerful system for machine learning and classification and comes along with a high number of predefined functions. The TextTrend application plugins apply these functions by invoking a general TextTrend WEKA plugin.

The **R** library plugin is used via **R** scripts accessible from application plugins, such as the one provided by the Dynnet package. Most of TextTrend's other **R** scripts are located in a separate, user-extendable bundle (so that every user or developer can add her own **R** scripts that can be run from the CiShell/TextTrend menu).¹

A key feature of the TextTrend framework is “trending”, which means the production of structured representations along a selected dimension, considered a “trend parameter”, which is typically (but not exclusively) time. Using trending, we can produce time-stamped (or, in general, trend parameter stamped) sets of networks from textual or other, structured or unstructured sources of data. For example, from a corpus of scientific publications we may

¹ Currently the entire **R** environment is needed to run the **R** scripts in the TextTrend system, but work is under way to migrate this to JRI (part of rJava project; <http://rosuda.org/rJava>), which is an interface for calling **R** codes from Java.

extract various matrices (similar to document-term matrices), the columns or rows of which encode individual attributes (e.g. keywords, title, authors, or year etc.); then, by applying the trending operator, one of the columns can be selected as a “trend parameter”. Using values of the trend parameter, we can generate a series of new matrices, one matrix for each of the selected individual trend attribute values (or intervals, e.g. years or decades, but also authors or author groups). Using this simple but powerful technique we can split every aggregate model into a series of instances that can be handled together, making it possible to introduce the analysis techniques such as discussed in Section 2 of this paper.

3.4. Integrated algorithms and class libraries

The CIShell framework can run several types of algorithms, such as converters, validators, algorithms implemented in Java, as well as native algorithms, which can be any executable programs wrapped in the proper way. The OSGi approach makes it possible that each usch algorithm is implemented as a bundle. Also, a bundle can be a class library, so other algorithms can use it as a “shared lib”. The following main algorithms are integrated in TexTrend at the moment:

- **org.texttrend.textprocessing.lib.uima** - UIMA class library (<http://incubator.apache.org/uima/>),
- **org.texttrend.classification.lib.weka** - WEKA class library, (<http://www.cs.waikato.ac.nz/ml/weka/>),
- **org.texttrend.statistics.R** - The R Project for Statistical Computing (<http://www.r-project.org/>),
- **org.texttrend.visualization.cytoscape** - Network analyzer and visualizer (<http://www.cytoscape.org/>),
- **org.texttrend.converter.arff2csv** - ARFF2CSV converter,
- **org.texttrend.converter.csv2arff** - CSV2ARFF converter,
- **org.texttrend.converter.arff2xgmml** - ARFF2XGMML converter,
- **org.texttrend.data** - implementations of accessible data types in TexTrend,
- **org.texttrend.dataloader** – loader of data types (currently only the CSV loader is implemented)

As we write this, TexTrend has the following data implementations:

- Attribute-Relation File Format (<http://www.cs.waikato.ac.nz/~ml/weka/arff.html>), file based,
- Comma Separated Values (http://en.wikipedia.org/wiki/Comma-separated_values), file based,
- eXtensible Graph Markup and Modeling Language (<http://en.wikipedia.org/wiki/XGMML>), file based.

3.5. Composability, interoperability, extendability

The TexTrend system is work in progress, pursued as a project funded for 3 years. It is currently in its first implementation phase, with several extensions to come soon (in particular, functional modules, but also class libraries and various new services such as a web service using the workflow functionality are under development). Numerous extensions will concern future dynamic network analysis modules and other uses of the trending operation.

The fully modular plugin based service architecture of the underlying CIShell/OSGi layer reduces all this development to a matter of adding new bundles to achieve automatic interoperability with the core of TexTrend and with virtually any other, similar systems, such as the aforementioned NWB, or Cytoscape (v3.0) which is recently using OSGi, etc. etc.

4. Application example

4.1. SNA, science mapping and dynamic networks

A prominent consumer of the techniques and concepts yielded from social network analysis is scientometrics. Network analysis provides the central framework to map and measure the organization of science and technology since the emergence of the field, and its applications cover several dimensions of the subject matter.

The original domain of SNA is instantiated in the analysis of social structures emerging in S&T: these typically include patterns of research collaboration in different levels of aggregation. Features of co-author networks are used as proxies for arguing in the sociology of science, while graphs of institutional research collaborations are utilized in testing mainstream hypotheses in science and technology policy, such as the degree of academia-industry relationships, or the status of regional collaborations.

The most salient stream of scientometrics, citation analysis (or bibliometrics in its narrow sense) also relies on the network approach. Based on the use of citation information, various aspects of scholarly communication are modeled by citation networks. Connecting the citing publication with its references, i.e. inter-citation studies, result in social networks with various actors (authors, journals, etc.), conveying sociological or historiographic content. Formation of scientific communities and the spread of ideas exemplify these aspects. On the other hand, constructing nets from citation co-occurrences represents the intellectual background of a given piece or trend of research.

Furthermore, SNA finds its uses in more abstract networks common to science mapping exercises. The now-classical method of revealing emerging research trends via infoscience means the construction of co-word maps. Co-word maps are constructed from thematic indicators (keywords, subject category assignments, title words, etc.) based on their co-occurrence in document descriptions. The resulting network is well suited for the knowledge discovery task of revealing trendsetting topics or concepts that organize the scientific discourse as well as distinctive contexts for a specific topic, and, most interestingly, topical clusters that constitute the thematic map of the field under study.

In co-word analysis it is consensual to deploy the measure of betweenness centrality of nodes (words) that accounts for the relative importance of nodes in a particular network (Chen 2004, 2006, also cf. Leydesdorff, 2007). Importance is interpreted here as the capacity of the node to connect other (groups) of nodes to each other. Though this concept could well be described via pure graph theoretical means, the meaning of the measure is clearly parallel to the social interpretations, from where it was borrowed. Just as social actors that connect otherwise unconnected groups (high betweenness centrality) can organize the community, so the concepts that connect distinct contexts can organize the discourse and create links and transitions from one context to the other (transient pattern, Chen, 2006).

Table 1 collects typical types of networks common to science mapping, along with their analytic concepts borrowed from SNA.

Dimension	Networks under study	Characteristic concept(s) from SNA
Social	Networks of collaboration (authors, institutions etc.)	Degree distribution, community detection, centrality measures
Social, intellectual, historical	Citation networks (inter-citation, co-citation etc.)	Information flow, path length
Thematic, semantic	Co-word maps (keyword, subject area etc.)	Clusters, centrality measures

Table 1.

The need to understand trends and developments in these dimensions naturally calls for dynamic network analysis as put forward in this paper. Although it is intuitively clear, it turns out to be a challenging task to formally determine how changes should be captured and characterized in dynamic networks to represent the trends and meaningful patterns. Structural change on the encompassing network level lends itself to a quantitative analysis more straightforwardly, as overall measures like parameters of degree distribution are easily expressed as functions of the trend parameter (i.e. time). However, when it comes to tracking the evolution of network-related properties of nodes or groups of nodes (e.g. importance of concepts over time), or even the evolution of meaningful clusters (e.g. development of groups of related concepts), the traditional measures prove to be insufficient for the task.

In what follows, we demonstrate a scenario in scientometrics, where the solutions proposed by the TexTrend framework provide a cutting edge advantage. This example extracts information immanent to the dynamic aspect of scientometric networks via both visualization and quantitative analysis.

4.2. Example: Dynamic author networks on „academic career”

We conducted an analysis on a community defined by the research topic *academic career*. We have retrieved a full set of bibliographic records from the ISI Web of Science database controlling for the phrase „academic career” being present in either the title or the abstract, or among the keywords of the item. The resulting dataset consisted of approximately 400 publication records covering a 35-years period from 1975 to 2009.

4.2.1. Visualization.

To reveal the most characteristic community patterns behind the topic, the corpus was subjected to co-author analysis. Since our primary interest was to contrast two perspectives, the static and the dynamic view on collaboration structures, the co-author networks have been obtained from the data in two phases. First, the weighted graph of common authorships was extracted from the entire dataset as a whole, yielding an all-in-one snapshot of the pattern, which realizes a static view of the community (not shown). At the second phase, bibliographic data have been split into six (equal) time periods, and networks were extracted for each of the resulting intervals. The latter approach resulted in a time series of networks indexed by the trend parameter, time.

This graph family was subsequently visualized in an iterative manner. In each iteration, a filter acting on component size has improved the clarity of the structure: at first, each node

was plotted, many of which are isolates and not participating in the network. In each iteration, connected components were plotted that exceeded a threshold level in terms of size, therefore representing real subgroups in the net. Figure 7 and Figure 8 show the two extremes of this process. On Figure 7, all actors are visible, while in Fig. 8 only those components that fall above the 3rd quartile of the component size distribution for the respective network (time slice) are shown.

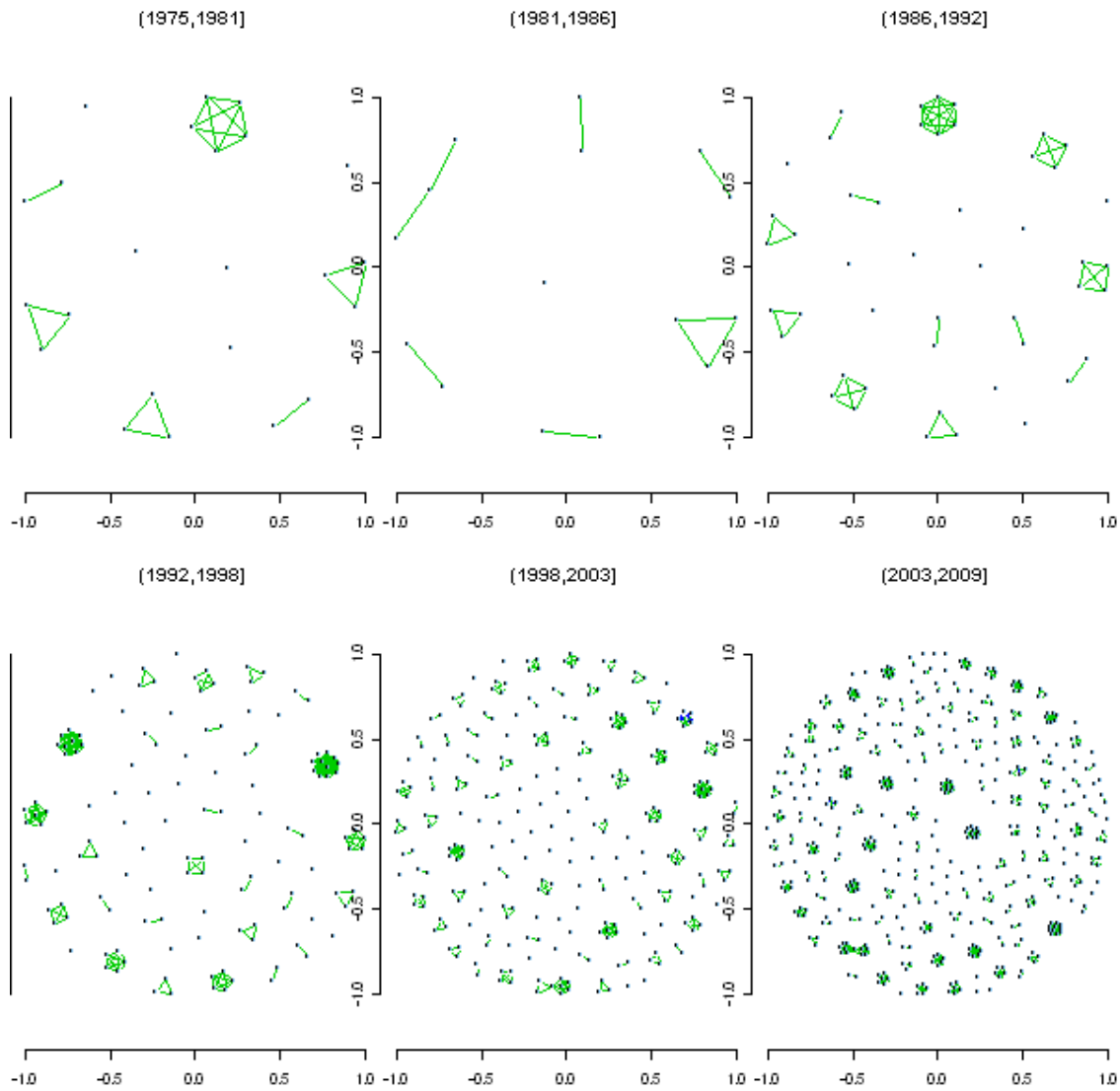


Figure 7. Raw data representation

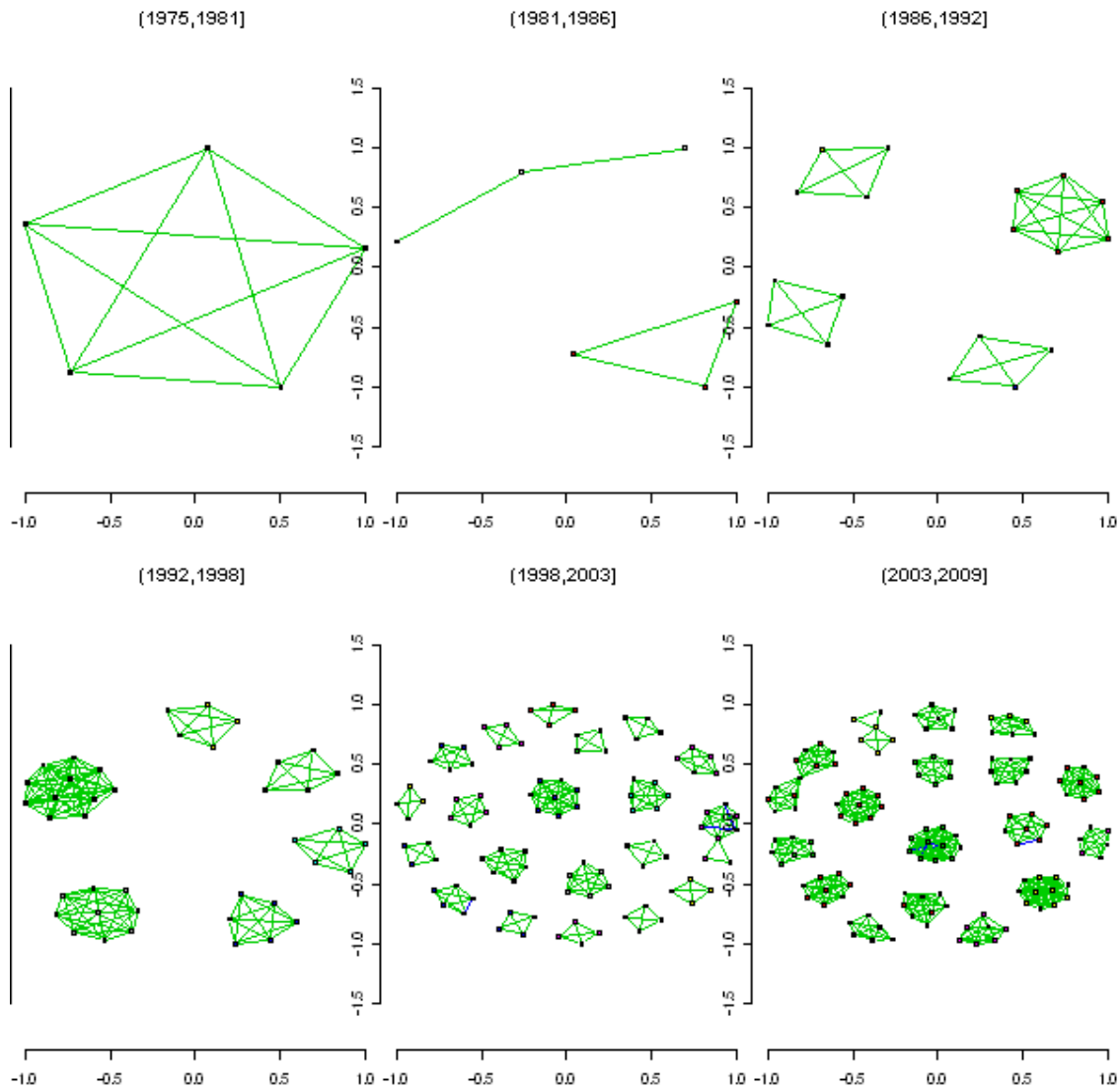


Figure 8. Filtered representation using component size as a threshold

4.2.2. Analysis of the networks.

The main virtue in the visualizations of network dynamics is that they provide the capacity to the analyst to discover several trends by pure visual inspection. Fig. 7 makes it clear that the number of authors, and, more importantly, that of single authors has been significantly increasing since 1975. Also observable in Fig 7, community formation (connected components) has, from the beginning, been restricted to the emergence of small and dense subgroups, that are typically attributable to single multi-authored publications rather than to systematic and well-developed collaborations. This hypothesis is further strengthened by Fig. 8, exposing the largest groups in each period. The tendency of expansion both in terms of the

number and cardinality of multi-authorship groups becomes salient on the snapshots, while a small differentiation in the structure of these groups is being present for the last two intervals. Among the fully connected subgraphs we can find two communities in (2003,2009] that contain a central author, each connecting two dense groups. Characterized by a high value of betweenness centrality, these authors organize a higher-order community, because those, among other things, cover more than one publication as the basis of coherence.

As contrasted to a usual static snapshot of the network extracted from the corpus as a whole, the dynamic representation evidently provides much deeper insight into the very process of network formation. Omitting the information on the synchronous co-existence of groups suppresses, among others, the fact that publishing in broad alliances is a relatively late phenomenon in this domain of research, while becomes predominant in the second half of the whole window under study.

Central to the conception of the TexTrend framework is to provide, beyond visualization, well-developed and well-chosen quantitative measures in order to both confirm and extend the results from visual inspection. To capture the tendencies described above, we applied four structural indicators designed for the characterization of either the network or the node level (Figure 9). We then compiled a time series of these indicators, mirroring the periodical changes in their values. Applied measures included the number of connected components having at least 3 nodes (#clus), the average weight of the links in these components (mean.weight), the average density of the subnets (mean.dens), and also the mean of the average betweenness centrality of the nodes calculated for each component (mean.bw).

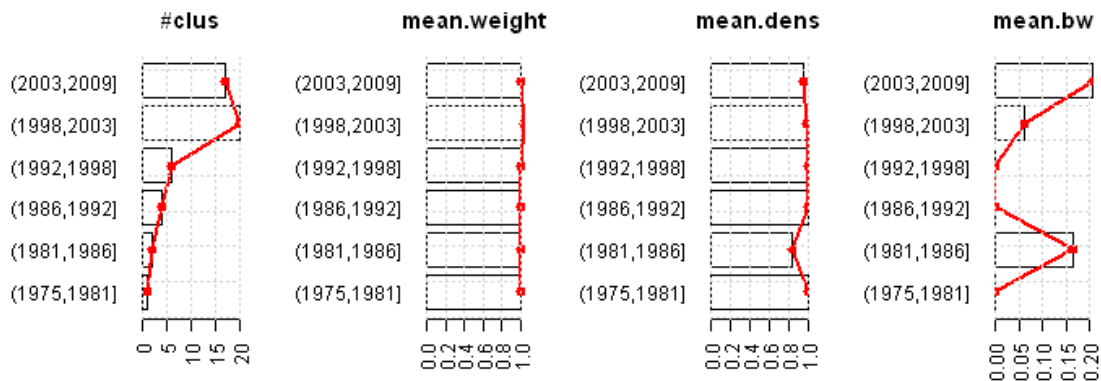


Figure 9. Quantitative analysis of an example dynamic co-author network

Results, as showed in Figure 9, confirm and also summarize our previous characterizations of the changes. The number of connected groups shows a sudden increase and high values for the last two periods. The more-or-less uniform weight and density values, being almost invariably 1 in each interval (though this value means an absolute weight and a relative density number), clearly demonstrate the nature of these groups as being usually formed upon a single publication. Finally, but most strikingly, mean betweenness values for each time slice perfectly indicate the emergence of central actors in the last two periods (with an outlier in the second time slice, which is due to a small „pseudo-group” with three sequentially connected

members; this outlier can be filtered out both by the visual inspection, and by applying an additional measure for the distribution of mean component sizes over time, not done here). The integration of the four structural indicators as well as the flip book representation using **R** scripts is under way in the TextTrend system.

5. Summary and Conclusions

Dynamic networks offer significant new challenges and correspond to a largely untrodden ground as even the basic tools and concepts are missing. Yet the analysis of dynamic networks is a significant task motivated by challenges from many real-world problems that invite several new approaches. In this paper we have discussed a set of new measures and their integration into a toolkit that uses a flexible, integrative system, the TextTrend system, based on the OSGi driven CShell platform. The new measures treat dynamic network evolution at both global and local, as well as cluster level, and come along with a set of visualization tools in the TextTrend system. We demonstrated the use of these concepts and tools on a few SNA/bibliometrics examples.

Acknowledgments

This research was partly supported by the Hungarian Government (Anyos Jedlik programme managed by the National Office for Research and Technology, <http://www.nkth.gov.hu/>) through the TextTrend project (www.texttrend.org), contract no. NKFP_07_A2 (2007)/TEXTREND and the European Union through the DynaNets project (www.dynanets.org), EU project no. FET-233847. The supports are gratefully acknowledged.

References

- Barabasi, A-L, and Albert, R. 1999: Emergence of Scaling in Random Networks, *Science* 286, pp. 5439 -550.
- Barabasi, A-L. 2009: Scale-Free Networks: A Decade and Beyond, *Science* 325, p. 412.
- Barrat, A., Barthelemy, M. Vespignani, A. 2008: Dynamical processes in complex networks *Cambridge University Press*.
- Bender-deMoll, S., McFarland, D.A. 2006: The Art and Science of Dynamic Network Visualization. *Journal of Social Structure*. Volume 7, Number 2.
- Börner, K., Chen, Ch., Boyack, K.W. 2003: Visualizing Knowledge Domains, In: Blaise Cronin (Ed.), *Annual Review of Information Science & Technology (ARIST)* (Vol. 37, pp. 179-255). Medford, NJ: Information Today, Inc./American Society for Information Science and Technology.
- Börner, K., Herr, B.W.II, Fekete, J-D. 2007: IV07 Software Infrastructures Workshop. *Paper presented at the 11th International Conference on Information Visualization*, Zurich, Switzerland.
- Börner, K. 2009: Plug-and-Play Macroscopes, *Comm. ACM*, in press.

- Carley, K. 2003: Dynamic Network Analysis. 133-145. Committee on Human Factors, National Research Council.
- Carley, K. 2004: <http://privacy.cs.cmu.edu/dataprivacy/papers/socialnetworks/index.html>
- Carley, K., Diesner, J., Reminga, J. and Tsvetov, M. 2005: Toward an Interoperable Dynamic Network Analysis Toolkit. *DSS Special Issue on Cyberinfrastructure for Homeland Security*.
- Chen, C. 2006: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377.
- Chen, C. 2004: Searching for intellectual turning points: Progressive Knowledge Domain Visualization. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), 101 (Suppl. 1), 5303-5310.
- Cyberinfrastructure for Network Science Center, Indiana University, Bloomington (2008). Cyberinfrastructure Shell: Core Specification 1.0. <http://cishell.org/dev/docs/spec/cishell-spec-1.0>
- Dynnet 2009: Documentation of the Dynnet package, www.textrend.org/Publications/Dynnet_Doc_v1.0.pdf
- Falkowski, T., Bartelheimer, J., Spiliopoulou, M. 2006: Mining and visualizing the evolution of subgroups in social networks, In: *2006 IEEE/WIC/ACM International Conference on Web Intelligence* (WI 2006 main conference proceedings),
- Ganguly, N.; Deutsch, A., Mukherjee, A. (Eds.) 2009: *Dynamics On and Of Complex Networks. Applications to Biology, Computer Science, and the Social Sciences*. Series: Modeling and Simulation in Science, Engineering and Technology, Birkhauser, New York.
- Jordán, F., Scheuring, I., 2004. Network Ecology: topological constraints on ecosystems dynamics. *Physics of Life Reviews* 1:139-172
- Lahiri, M. Berger-Wolf, T.Y. 2008: Mining Periodic Behavior in Dynamic Social Networks, *Eighth IEEE International Conference on Data Mining*, ICDM '08, pp. 373-382.
- Leskovec, J. 2008. *Dynamics of large networks*, Ph.D. Dissertation, Machine Learning Department, School of Computer Science, Carnegie Mellon University, Technical report CMU-ML-08-111, September
- Leydesdorff, L. 2007: Betweenness centrality as an indicator of the interdisciplinary of scientific journals. *Journal of the American Society for Information Science and Technology* 58 (9), 1303-1319.
- Moody, J., McFarland, D.A., and Skye Bender-deMoll 2005: Visualizing Network Dynamics, *American Journal of Sociology*
- Pennock, D. M., G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles 2002: Winners Don't Take All: Characterizing the Competition for Links on the Web. *PNAS* 99, 5207-5211.
- Shalizi, C., Camperi, M.F., and Klinkner, K.L. 2007: Discovering Functional Communities in Dynamical Networks, in: Anna Goldenberg, Edo Airoldi, Stephen E. Fienberg, Alice Zheng, David M. Blei and Eric P. Xing (eds.), *Statistical Network Analysis: Models, Issues and New Directions*, New York: Springer-Verlag, pp. 140—157.
- Sloot, P.M. A., Ivanov, S.V., Boukhanovsky, A., van de Vijver, D.A., Boucher, C. A. B. 2008: Stochastic simulation of HIV population dynamics through complex network modeling, *Int. J. Comput. Math.* 85(8): 1175-1187

Sloot, P.M.A., Peter V. Coveney, G. Ertaylan, V. Müller, C.A. Boucher and M. Bubak 2009: HIV decision support: from molecule to man, *Phil. Trans. R. Soc. A* **367**, 2691-2703.

Sonia 2009:

<http://www.stanford.edu/group/sonia/dataSources/index.html>.

Szakolczi, Z. 2009: A tudományometriai vizsgálatok során előálló, időben (vagy egyéb dimenzió mentén) változó hálózat-sorozatok elemzése, MSc. Thesis (in Hungarian)

Trier, M., Bobrik, A. 2007: Analyzing the Dynamics of Community Formation using Brokering Activities, In: *Proceedings of the Third Communities and Technologies Conference*, Michigan 2007, Springer Series.

Wikipedia 2009:

http://en.wikipedia.org/wiki/Social_network_analysis_software