Spectral Network Analysis

► Spectral Analysis

Spectral Technique

► Spectral Analysis

Stability and Evolution of Scientific Networks

Eugenia Galeota[1], Susanna Liberti[1],
Frederic Amblard[2], and Walter Quattrociocchi[3,4]
[1]Department of Biology, University of Rome Tor Vergata, Rome, Italy
[2]IRIT – Universit Toulouse 1 Capitole, Toulouse Cedex 9, France
[3]London Institute of Mathematical Sciences, London, UK
[4]Natwork Department, IMT Alti Studi Lucca, Lucca, Italy

Synonyms

Emergence; Evolving structures; Scientific communities; Social network analysis; Social selection; Temporal metrics; Time-varying graphs

Glossary

APS American Physical Society
DOI Digital Object unique Identifiers associated to papers
Egalitarian Growth The growth benefiting on average equally to each node
Evolution of Social Networks The change in time of the structure of a social network due to changing interactions between the components
TVG Time-Varying Graphs

Introduction

Nowadays one of the most pressing as well as interesting scientific challenges deals with the analysis and the understanding of social systems' dynamics and how these evolve according to the interactions among their components. The efforts in this area strive to understand what are the driving forces behind the evolution of social networks and how they are articulated together with social dynamics – e.g., opinion dynamics, the epidemic or innovation diffusion, the teams formation, and so forth (Deffuant et al. 2001; Moore and Newman 2000; Lelarge 2008; Butts and Carley 2007; Powell et al. 2005; Guimera et al. 2005; Quattrociocchi et al. 2009, 2010). In this paper we approach the challenge of depicting the evolution of social systems from a network science's perspective. As an example, we chose the case of scientific communities by analyzing a portion of the American Physical Society dataset (APS). The analysis addresses the coexistence of coauthorship and citation behaviors of scientists. On the one hand, the studies on scientific network dynamics deal with the understanding of the factors that play a significant role in their evolution, not all of them being neither objective nor rational, e.g., the existence of a star system (Wagner and Leydesdorff 2005; Newman 2001a, 2004a; Jeong et al. 2002), the blind imitation concerning the citations (MacRoberts and MacRoberts 1996), and the reputation and community affiliation bias (Gilbert 1977). On the other hand, having some elements to understand such dynamics could enable for a better detection of the hot topics and of the vivid subfields and how the scientific production is advanced with respect to the selection process inside the community itself. Among the available data to analyze such a system, a subset of the publications in a given field is the most frequently used such as in De Solla Price (1965), Newman (2001b); **?**, Quattrociocchi et al. (2012), Amblard et al. (2011), Santoro et al. (2011), and Radicchi et al. (2009). The scientific publications correspond to the production of such a system and clearly identify who are the producers (the authors), which institution they belong to (the affiliation), which funded project they are working on (the acknowledgement), and what are the related publications (the citations), having most of the time a public access to these data explain also a part of its frequent use in the analyses of the scientific field. Classical analyses concern either the coauthorships network (Jeong et al. 2002; Newman 2001a) or the citation network (Hummon and Dereian 1989; Redner 2005), more rarely the institutional network (Powell et al. 2005). Moreover, such networks are often considered as static and their structure is rarely analyzed over time (an exception is the one performed by Radicchi et al. (2009), and Leskovec et al. (2005a). The illustrative analysis presented in the paper passes through different data transformations aimed at providing different perspectives on the APS network and its evolution. In Newman (2001a) the network of scientific collaborations, explored upon several databases, shows a clustered and small-world structure Watts (1999) and Tang et al. (2009). Moreover, several differences between the collaborations' patterns of the different fields studied are captured. Such differences have been deepened in Newman (2004a) with respect to the number of papers produced by a given group of authors, the number of collaborations, and the topological distances between scientists. Peltomaki and Alava in Peltomaki and Alava (2006) propose a new emulative model aimed at approximating the growth of scientific networks by incorporating bipartition and sublinear preferential attachment. A model for the self-assembly of creative teams based on three parameters (e.g., team size, the rate of newcomers in the scientific production, and the tendency of authors to collaborate with the same group) has been outlined in Guimera et al. (2005). Connectivity patterns in a citations network have been studied with respect to the development of the DNA theory (Hummon and Dereian 1989). The work of Klemm and Eguiluz (2002) observed that real network (e.g., movie actors, coauthorship in science, and word synonyms) growing patterns are characterized by a clustering trend that reaches an asymptotic value larger than regular lattices of the same average connectivity. In the field of social network analysis, several works have approached the problem of temporal metrics (Holme 2005; Kostakos 2009; Kossinets et al. 2008). The focus is on the definition of instruments able to capture the intrinsic properties of complex systems' evolution, that is, characterizing the interdependencies and the coexistence between local behaviors (interactions) and their global effects (emergence) (Davidsen et al. 2002; Mataric 1992; Woolley 1994; Deffuant et al. 2001; Quattrociocchi et al. 2010). The research approach to characterize the evolution patterns of social networks at the very beginning was mainly based upon simulations, while in the past few years, due to the large availability of real datasets, either the methodology of analysis or the object of research has changed (Taramasco et al. 2010; Leskovec et al. 2007; Kossinets et al. 2008).

Analysis of Scientific Network Dynamics

In this work we present a very basic analysis aiming at understanding the social aspects of the scientific systems by coupling the collaborations between scientists and their effect on the scientific community itself through the citation network. The data to build up the networks analyzed in this work has been extracted from the APS (American Physical Society) dataset, made available upon request by the APS for research purposes. The database contains information about 463,343 articles published on 11 journals of the APS in a time span ranging from 1892 to 2009. For the citations network we used a list of 2,944,144 DOI pairs in which the first DOI identifies an article containing a reference to the article identified by the second DOI. A date flag corresponding to the issue date of the citing article has been associated to each couple of DOIs in the list to represent the citation date. Such information has been obtained from the "Article metadata" part of the database which is divided by journal and provides for each paper the following fields: DOI, journal, volume, issue, first page, and last page OR article ID and number of pages, title, authors, affiliations, publication history, PACS codes, table of contents heading, article type, and copyright information. The list of authors provided for each DOI has been used to generate the collaboration network where authors of the same paper form small coauthorship cliques. Starting from the metadata a list of 17,069,841 total coauthorships has been generated for 119,172 unique authors' surnames. In order to assign a date to the collaboration, the submission date of the coauthored article has been associated to each couple of authors. The data transformation is performed through the *time-varying graphs* formalism. The *time-varying graph* (TVG) formalism, recently introduced in Casteigts et al. (2010), is a graph formalism based on an *interaction-centric* point of view and offers concise and elegant formulation of temporal concepts and properties (Santoro et al. 2010). Let us consider a set of entities V (or *nodes*), a set of relations E among entities (*edges*), and an alphabet L labeling any property of a relation (*label*), that is, $E \subseteq V \times V \times L$. The set E enables multiple relations between any given pair of entities, as long as these relations have different properties, that is, for any $e_1 = (x_1, y_1, \lambda_1) \in E$, $e_2 = (x_2, y_2, \lambda_2) \in E$, $(x_1 = x_2 \wedge y_1 = y_2 \wedge \lambda_1 = \lambda_2) \implies e_1 = e_2$. Relationships between entities are assumed to occur over a time span $\mathbb{T} \subseteq \mathbb{T}$, namely, the *lifetime* of the system. The temporal domain $\mathbb{T}$ is assumed to be $\mathbb{N}$ for discrete-time systems or $\mathbb{R}$ for continuous-time systems. The time-varying graph structure is denoted by the set $\mathcal{G} = (V, E, \mathbb{T}, \rho, \zeta)$, where $\rho : E \times \mathbb{T} \to \{0, 1\}$, called *presence function*, indicates whether a given edge is present at a given time and $\zeta : E \times \mathbb{T} \to \mathbb{T}$, called *latency function*, indicates the time it takes to cross a given edge if starting at a given date. As in this paper the focus is on the temporal and structural analysis of a social network, we will deliberately omit the latency function and consider TVGs described as $\mathcal{G} = (V, E, \mathbb{T}, \rho)$. Given a TVG $\mathcal{G} = (V, E, \mathbb{T}, \rho)$, one can define the *footprint* of this graph from t_1 to t_2 as the static graph $G^{[t_1,t_2)} = (V, E^{[t_1,t_2)})$ such that $\forall e \in E$, $e \in E^{[t_1,t_2)} \iff \exists t \in [t_1, t_2), \rho(e, t) = 1$. In other words, the footprint aggregates interactions over a given time window into static graphs. Let the lifetime $\mathbb{T}$ of the time-varying graph be partitioned in consecutive subintervals $\tau = [t_0, t_1), [t_1, t_2) \ldots [t_i, t_{i+1}), \ldots$, where each $[t_k, t_{k+1})$ can be noted τ_k. We call *sequence of footprints* of $\mathcal{G}$ according to τ the sequence $SF(\tau) = G^{\tau_0}, G^{\tau_1}, \ldots$.

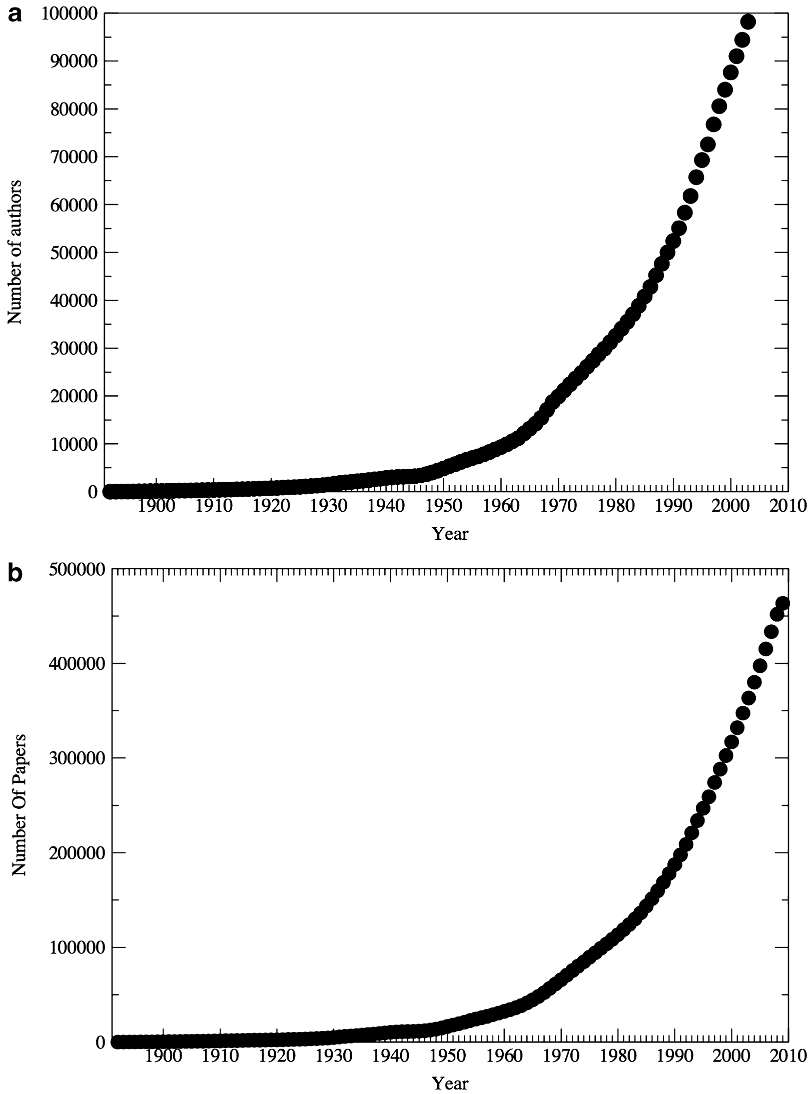
Hence, we derive two time-varying graphs: the *temporal coauthorships network*, with undirected edges and authors as nodes where a link stands for the relations of coauthoring a paper and the *temporal citations network* having papers as nodes and the links (directed) representing the citations from a paper to another one. The temporal dimension of both networks is derived by the paper's submission date. The temporal coauthorship network has edges labeled with the date of submission, while the temporal citations network has the nodes labeled with the publication date of papers citing other papers.

More formally, we can define this system as two networks:

- The **temporal coauthorships network** as a quadruplet $G_a^t : (V, E, \mathbb{T}, \rho)$, where the nodes in $v \in V$ are the authors and links $e \in E$ connect a couple of scientists coauthoring a paper. The temporal domain $\mathbb{T} = [t_a, t_b)$ of the function ρ is the *lifetime* of each node v that in this context is assumed as t_a to be the submission date of the paper and $t_b = \infty$.
- The **temporal citations network** as a quadruplet $G_c^t : (V, E, \mathbb{T}, \rho)$, where the nodes in the set V are the papers and each edge $e \in E$ corresponds to a citation to another paper. As for the coauthorships network, the temporal dimension $\mathbb{T} = [t_a, t_b)$ of the presence function ρ of G_c^t is defined within the submission date of papers and ∞.

Networks Evolution

In Fig. 1 we show the number of authors and the number of papers for each year. One can observe from such figures an exponential growth of both the number of authors and of papers along time. Such results are not surprising and have been highlighted by several former works (for instance in Radicchi et al. (2009)). The exponential growth

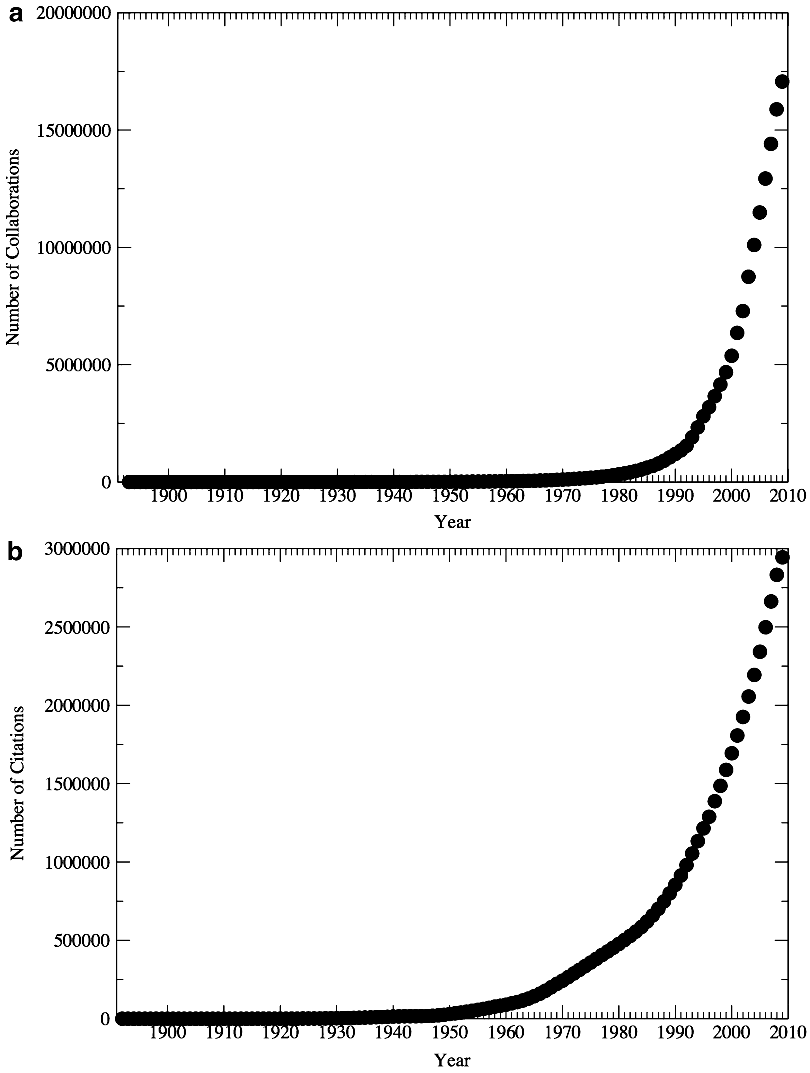


Stability and Evolution of Scientific Networks, Fig. 1 (**a**) Number of authors and (**b**) number of papers

in the number of publications is more or less directly attributed to a change in the behavior of scientists induced by the pressure to publish all along their career (it has been popularized through the proverb publish or perish). The exponential growth of the number of authors is more surprising at a first attempt, as it does not translate an exponential increase of the positions in research that does not exist. It is much more seriously explained by an indirect effect of the exponential growth of publications. We have to remind that this dataset concerns the APS publications, and such publications do not render the effective number of physicists. As the popularity of the APS journals increases, they probably attract more and more physicists worldwide, and we can expect a stabilization of such tendency once as the APS will tend to reference nearly the whole population of physicists worldwide.

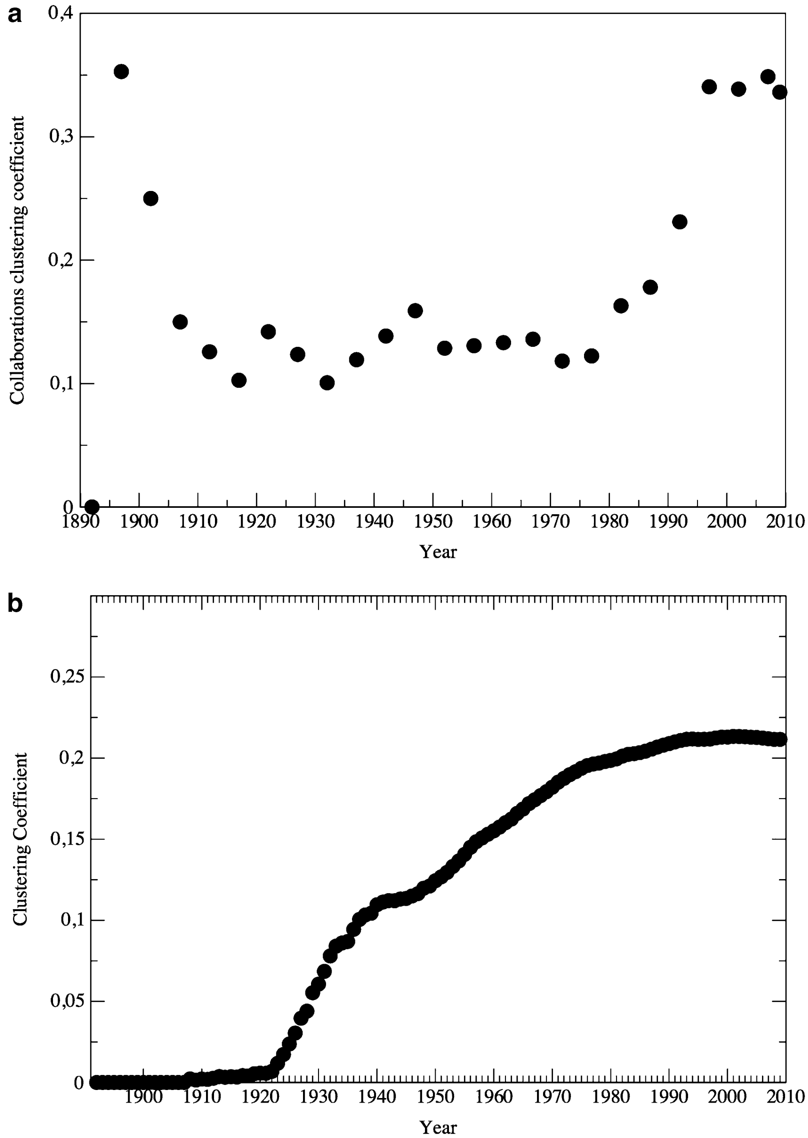
In Fig. 2 we show the number of collaborations within authors and the number of citations within papers. Those two measures correspond basically to the number of edges in each of the two networks. The first important element concerns the increase of the number of collaborations that scales as a power law rather than an exponential. This feature results clearly of a double effect over the past few years. The first one is

S



Stability and Evolution of Scientific Networks, Fig. 2 (**a**) Number of collaborations and (**b**) number of citations

directly linked to the increase in the number of papers that increases the potential of collaboration among authors. The other effect comes from the progressive increase of the number of authors per paper. Translated into network terms, it means that each paper coauthored by N scientist creates $N*(N-1)/2$ links in the collaboration network. As a consequence, if you follow the current tendency to increase the number of authors, you increase the power coefficient, the number of links among them. Concerning the other figure, the exponential growth is probably less essential but again it results from two combined effects. On the one hand, the number of papers published increases in the same way as the total number of citations. On the one end, the slight tendency to progressively increase the number of papers cited in each paper straightened again the slope. Considering the two graphs mentioned in Fig. 1, the basic feature that we can observe is a global tendency of the increase of the number of nodes in the corresponding networks. The point that the number of links on each graph increases more rapidly than the number of nodes leads to the conclusion that the coauthorship and the citation graphs tend to grow and densify as well. However, we don't have any clue concerning the properties of such a density growth, mainly, is it an egalitarian growth or is it an elitist system with some few nodes benefiting from this



Stability and Evolution of Scientific Networks, Fig. 3 The evolution of the clustering coefficient. (**a**) Collaboration network. (**b**) Citation network

increase in density and the majority of nodes being left behind without many links? The measure of the evolution of the clustering coefficient on such networks can bring arguments for this distinction.

In Fig. 3 we show the clustering coefficient – i.e., the transitivity among nodes – for the collaboration and citation network. Qualitatively, the curves are totally different on the two networks. On the coauthorship network, the evolution follows first an important decrease and then stabilizes before increasing again. On the citation network instead, we can observe an increase that tends to stabilize in the last 20 years. The elements of interpretation behind those two figures are the following. From the coauthorship network, the first global decrease can be explained mainly because it starts from an important number of non-connected components in the network. Therefore, the creation of new links among those components (or communities) that corresponds to a porosity of the different communities in physics results in a global decrease of the clustering coefficient as it tends to dissolve locally the density of each component. Once a global giant component is created (corresponding to the observe plate on the figure), then there is a stabilization of the clustering coefficient. The final increase is maybe the most interesting feature of this analysis as it corresponds to the case where, in a global single component, the clustering is increasing. This is the case where communities tend to emerge from a global network. Therefore this last increase could be interpreted as the formation and the radicalization of scientific communities on the global network. Such network communities correspond to the effective work in the scientific communities, i.e., coauthorship. Concerning the evolution of the clustering coefficient in the citation network, the first observation we have to make is that the global big component appears very soon on this network (this is much more probable to cite works from outside the field than to collaborate with people from outside the field). Therefore this global and progressive increase of the clustering coefficient corresponds solely to the progressive formation of scientific communities on the network. However, the final stabilization of the index results from a consolidation of the communities that have reached a relative equilibrium. We have to notice that in the case of the creation of new communities or emerging fields, we could see the global clustering coefficient increase again. Such an observation can be made on the figure where around 1940, we can observe a global stabilization of the index and therefore of the corresponding communities before than to increase again, such a new burst being the result of the inclusion of new communities in the network. However, in order to relativize such an effect, we have to remind that the dataset we analyze corresponds to the publications of the APS, and such an inclusion of new communities can result simply from an editorial choice corresponding to the launch of new journals on new thematics for the APS, but not necessarily for the scientific domain of physics.

Conclusions

In this paper we characterize the evolution of a scientific community extracted by the APS dataset. The temporal dimension and the metrics used for the analysis were formalized using time-varying graphs (TVG), a mathematical framework designed to represent the interactions and their evolution in dynamically changing environments.

Since we are interested in the relationships between collaborations and citation behaviors of scientists, we focus on the network of most cited authors and on its structural evolution where several interesting aspects emerge. Through our approach, we capture the role played by famous authors on coauthorship behaviors. They act as attractors on the community. The driving force is a sort of preferential attachment driven by the number of citations received by a given group that in terms of the goal of any scientific community indicates a strategy oriented to the community belonging.

Furthermore, the evolution of the network from a sparse and modular structure to a denser and homogeneous one can be interpreted as a threefold process reflecting the natural selection. The first phase is the exploration of ideas by means of separated works, once some ideas start to be cited (selected) more than others, then authors tend to join groups that have produced highly cited works. The selection is performed by individuals in a goal-oriented environment, and such a (social) selection produces self-organization because it is played by a group of individuals which act, compete, and collaborate in order to advance science. In fact, the driving force is an emergent effect of the interdependencies between citations and the goal of the scientific production since the social selection determines the emergence of a topic and of the scientists working on it by determining the so called preferential attachment toward groups and topics having high potential of citations.

Acknowledgments

Thanks to the American Physical Society for allowing us to use the APS dataset.

Cross-References

- ▸ Analysis and Visualization of Dynamic Networks
- ▸ Community Evolution
- ▸ Dynamic Community Detection
- ▸ Modeling and Analysis of Spatiotemporal Social Networks

References

Amblard F, Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2011) On the temporal analysis of scientific network evolution. In: CASoN, Salamanca, pp 169–174

Butts CT, Carley KM (2007) Structural change and homeostasis in organizations: a decision-theoretic approach. The Journal of Mathematical Sociology 31(4): 295–321

Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2010) Time-varying graphs and dynamic networks. Technical report, University of Carleton, Ottawa

Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a small world from local interactions: modeling acquaintance networks. Phys Rev Lett 88(12):128701

Deffuant G, Neau D, Amblard F, Weisbuch G (2001) Mixing beliefs among interacting agents. Adv Complex Syst 3:87–98

De Solla Price DJ (1965) Networks of scientific papers. Science 149(3683):510–515

Gilbert N (1977) Referencing as persuasion. Soc Stud Sci 7:113–122

Guimera R, Uzzi B, Spiro J, Amaral LA (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science 308(5722):697–702

Holme P (2005) Network reachability of real-world contact sequences. Phys Rev E 71(4):46119

Hummon NP, Dereian P (1989) Connectivity in a citation network: the development of DNA theory. Soc Netw 11(1):39–63

Jeong H, Neda Z, Ravasz E, Schubert A, Barabasi AL, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311:590–614

Klemm K, Eguíluz VM (2002) Highly clustered scale-free networks. Phys Rev E 65(3):036123+

Kossinets G, Kleinberg J, Watts D (2008) The structure of information pathways in a social communication network. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2008), Las Vegas, pp 435–443

Kostakos V (2009) Temporal graphs. Phys A Stat Mech Appl 388(6):1007–1023

Lelarge M (2008) Diffusion of innovations on random networks: understanding the chasm. In: Papadimitriou CH, Zhang S (eds) Internet and network economics. Springer, Berlin, Vol. 5385, pp 178–185

Leskovec J, Kleinberg J, Faloutsos C (2005a) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, Chicago. ACM, pp 177–187

Leskovec J, Kleinberg JM, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. TKDD 1(1)

MacRoberts MH, MacRoberts BR (1996) Problems of citation analysis. Scientometrics 36(3):435–444

Mataric M (1992) Designing emergent behaviors: from local interactions to collective intelligence. In: Proceedings of the international conference on simulation of adaptive behavior: from animals to animats, Honolulu, Hawai, USA. vol 2, pp 432–441

Moore C, Newman MEJ (2000) Epidemics and percolation in small-world networks. Phys Rev E 61:5678–5682

Newman MEJ (2001a) Proceedings of the National Academy of Sciences of the United States of America 98(2):404–409

Newman MEJ (2001b) Clustering and preferential attachment in growing networks. Phys Rev E 64:025102

Newman MEJ (2004a) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci 101:5200–5205

Newman MEJ (2004b) Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim E, Frauenfelder H, Toroczkai Z (eds) Complex Networks Lecture Notes in Physics Springer. Berlin/New York. Vol. 650, pp 337–370
Complex networks. Lecture notes in physics.

Peltomaki M, Alava M (2006) Correlations in bipartite collaboration networks. J Stat Mech 2006:P01010

Powell WW, White DR, Koput KW (2005) Network dynamics and field evolution: the growth of interorganizational collaboration in the life sciences. Am J Sociol 110(4):1132–1205

Quattrociocchi W, Paolucci M, Conte R (2009) On the effects of informational cheating on social evaluations: image and reputation through gossip. Int J Knowl Learn 5(5/6):457–471

Quattrociocchi W, Conte R, Lodi E (2010) Simulating opinion dynamics in heterogeneous communication systems. In: ECCS 2010, Lisbon

Quattrociocchi W, Amblard F, Galeota E (2012) Selection in scientific networks. Soc Netw Anal Min 2(3):229–237

Radicchi F, Fortunato S, Markiness B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80:056103

Redner S (2005) Citation statistics from 110 years of physical review. Phys Rev Phys Today 58:49–54

Santoro N, Quattrociocchi W, Flocchini P, Casteigts A, Amblard F (2010) Time varying graphs and social network analysis: temporal indicators and metrics. Technical report, University of Carleton, Ottawa

Santoro N, Quattrociocchi W, Flocchini P, Casteigts A, Amblard F (2011) Time-varying graphs and social network analysis: temporal indicators and metrics. In: 3rd AISB social networks and multiagent systems symposium (SNAMAS), York, Apr 2011, pp 32–38
Tang J, Scellato S, Musolesi M, Mascolo C, Latora V (2009) Small-world behavior in time-varying graphs. Arxiv preprint arXiv:0909.1712
Taramasco C, Cointet J-P, Roth C (2010) Academic team formation as evolving hypergraphs. Scientometrics 85:721–740
Wagner CS, Leydesdorff K (2005) Network structure, self-organization, and the growth of international collaboration in science. Res Policy 34(10):1608–1618
Watts DJ (1999) Networks, dynamics and the small world phenomenon. AJS 105:493–527
Woolley DR (1994) Plato: the emergence of online community. Comput Mediat Commun Mag 1(3):5

Statistical Analysis

► Extracting Individual and Group Behavior from Mobility Data

Statistical Inference

► Theory of Statistics, Basics, and Fundamentals

Statistical Modeling

► Siena: Statistical Modeling of Longitudinal Network Data

Statistical Models

► Theory of Statistics, Basics, and Fundamentals

Statistical Relational Learning

► Probabilistic Logic and Relational Models
► Relational Models

Statistical Relational Models

► Relational Models

Statistical Research in Networks – Looking Forward

Eric D. Kolaczyk
Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Synonyms

Propagation of uncertainty; Research challenges

Glossary

Network Summary Statistic A statistic summarizing a network graph

Propagation of Uncertainty Understanding the effect of uncertainty in an initial set of measurements on functions thereof

Introduction

The emerging field of network analysis, through its roots in social network analysis, has had a nontrivial statistical component from the start. In the ensuing years, problems in network analysis have motivated – and continue to motivate – new research in the field of statistics. Conversely, new developments in statistics are routinely integrated into network research. It is therefore rather surprising that, despite the many interesting and important statistical challenges in network analysis to which researchers have already been able to respond, there nevertheless are a number of challenges of an entirely fundamental nature that remain almost untouched!

We will support this central claim through two examples. Additional examples will be mentioned in passing at the end. All of these examples