Mapping change in large networks

M. Rosvall*

*Department of Biology, University of Washington, Seattle, WA 98195-1800**

C. T. Bergstrom†

*Department of Biology, University of Washington, Seattle, WA 98195-1800** *and*
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501‡

(Dated: July 13, 2009)

Change is a fundamental ingredient of interaction patterns in biology, technology, the economy, and science itself: Interactions within and between organisms change; transportation patterns by air, land, and sea all change; the global financial flow changes; and the frontiers of scientific research change. Networks and clustering methods have become important tools to comprehend instances of these large-scale structures, but without methods to distinguish between real trends and noisy data, these approaches are not useful for studying how networks change. Only if we can assign significance to the partition of single networks can we distinguish meaningful structural changes from random fluctuations. Here we show that bootstrap resampling accompanied by significance clustering provides a solution to this problem. To be able to connect changing structures with the changing function of networks, we highlight and summarize the significant structural changes with alluvial diagrams and realize de Solla Price's vision of mapping change in science: studying the citation pattern between about 7000 scientific journals over the past decade, we find that neuroscience has transformed from an interdisciplinary specialty to a mature and stand-alone discipline.

Researchers have developed a suite of network mapping tools to highlight important features while simplifying the overall structure of social and biological systems (1–6). With such tools we can abstract, quantify, and comprehend the nature of systems with numerous and diverse interacting components. As powerful as these tools have proven to be for understanding a system's structure, we do not yet have an adequate tool for mapping how this structure *changes*. For example: How has the network of global air traffic changed over the past half century? How does the organization of social contacts change when diseases develop and spread? How does the network structure of the federal funds market change when credit markets freeze up? How do gene regulatory networks differ between cancer and non-cancer states? And how does science itself evolve as paradigms shift through time? To quantify change in large networks, we must first identify the important structures, then assess what structures are statistically significant, and finally capture how these structures change.

Any tool for analyzing change must distinguish between meaningful trends and statistical noise. For example, statistical network models and stratified data make it possible to estimate global properties from the observation of sample networks (7–9). But when we are interested in the unique identities of the individual network components — Chicago O'Hare plays a unique and irreplaceable role in the global air traffic network, for example — we need another approach. Recent network approaches have become prominent in the study of complex systems because they can capture and respect the identities and characteristics of the components (10, 11). Often these individual differences matter critically and clustering rather than stratification must be employed to comprehend the data (1–6).

Moreover, many of the systems to which we apply network approaches are idiosyncratic in nature and preclude replicate observations. For example, there is one and only one global air traffic network for the year 2009. Therefore we cannot establish statistical significance by looking at multiple samples. Nor can we rely on temporal stability. While structures that remain unchanging over time may be statistically significant, we will not find significant changes by looking for features that stay the same.

One possibility would be to use a resampling technique such as the bootstrap, which assesses the accuracy of an estimate by resampling from the empirical distribution of observation (12). But we have only a single observation, a single network — so from what can we resample? When the single observation is composed of numerous components, as a network is composed of nodes and links, we can use the parametric bootstrap to assemble bootstrap networks by resampling from the components. Instead of resampling directly from the empirical distribution, a parametric model is used to fit the data. For the networks discussed in this paper, resampling nodes is not the right approach — it makes no sense to talk about the US air transit network without O'Hare, let alone the US network with two O'Hares. However, the link weights, which effectively define the nodes, can be parametrized and resampled without undermining the individual char-

*Electronic address: rosvall@u.washington.edu
†Electronic address: cbergst@u.washington.edu
‡URL: http://octavia.zoology.washington.edu/

acteristics of the nodes. With this approach we can assess the significance of clusters and estimate the accuracy of summary statistics, based on the proportion of bootstrap networks that support the observation (see Fig. 1).

Finally, to reveal stories in the network data and to be able to connect structural and functional changes, we use *alluvial diagrams* to highlight and summarize the significant structural changes. Our method could be applied to study, for example, how the global flight traffic pattern changes over time or how the federal funds market adapts structurally to cope with disturbances, but here we illustrate the method by mapping change in the structure of science itself (13).

Science is a dynamic, organized, and massively parallel human endeavor to discover, explain, and predict the nature of the physical world. In science, new ideas are built upon old ideas. Through cumulative cycles of modeling and experimentation, scientific research undergoes constant change: fields grow and shrink, merge and split. Citation patterns among scientific journals allow us to track this flow of ideas and how the flow of ideas changes over time (13). Here we use citation data from Thomson-Reuters' Journal Citation Reports 1997–2007, which aggregate, at the journal level, approximately 35,000,000 citations from more than 7000 journals over the past decade (14).

We first cluster the networks with the information-theoretic clustering method presented in ref. (5), which can reveal regularities of information flow across directed and weighted networks. We emphasize that, with appropriate modifications, our method of bootstrap resampling accompanied by significance clustering is general and works for any type of network and any clustering algorithm (see supplement for a detailed description of the method). To assess the accuracy of a clustering, we resample a large number $B \approx 1000$ of bootstrap networks from the original network. For the directed and weighted citation network of science, in which journals correspond to nodes and citations to directed and weighted links, we treat the citations as independent events and resample the weight of each link from a Poisson distribution with the link weight in the original network as mean (15) (see, for example, refs. (16–18) for other resampling techniques). Figure 1 illustrates an example network, the clustering of this network, and the clusterings of four of its bootstrap networks. For scalar summary statistics, it is straightforward to assign a 95% bootstrap confidence interval as spanning the 2.5th and 97.5th percentiles of the bootstrap distribution (18), but working with sets and assessing the accuracy of the clusters requires a different approach.

To identify the journals that are significant in their cluster assignments, we use simulated annealing to search for the largest significant subset of journals within each cluster of the original network that are clustered together in at least 95% of all bootstrap networks. To identify the clusters that are significantly distinct from all other clusters, we search for clusters whose significant subset

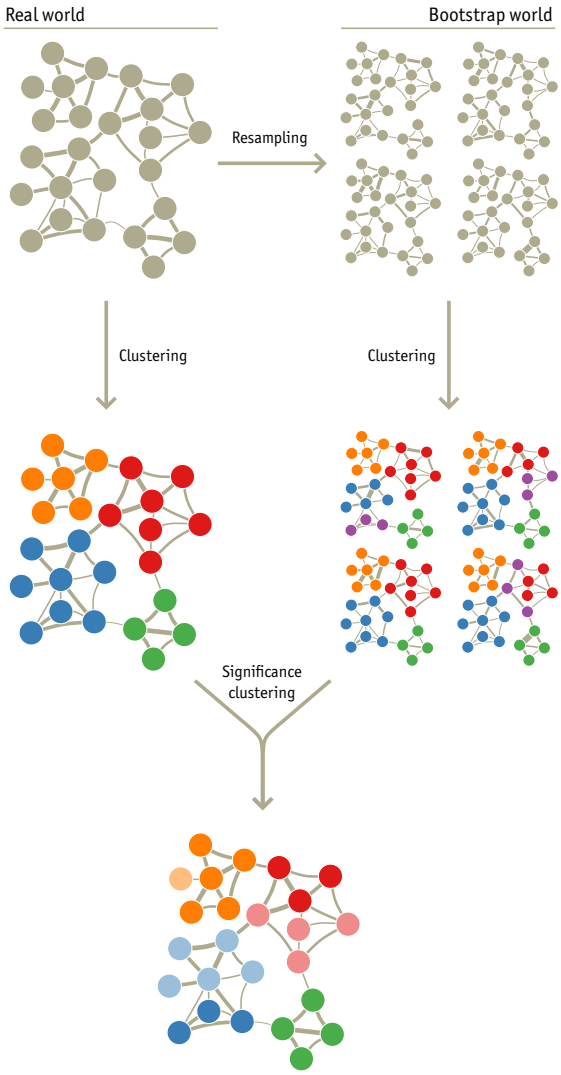

FIG. 1 Significance clustering of networks. The standard approach to cluster networks is to minimize an objective function over possible partitions of the network, as in the left side of the diagram. By repeated resampling of the weighted links from the original network, we create a "bootstrap world" of resampled networks. By clustering these as well, and comparing to the clustering of the original network, we can estimate the degree of support that the data provide in assigning each node to a cluster. In the bottom network, the darker nodes are clustered together in at least 95% of the 1000 bootstrap networks.

is clustered with no other cluster's significant subset in at least 95% of all bootstrap networks (see supplement). The significance-clustering step of Fig. 2 illustrates this process as applied to a network at two different time points.

Once we have a significance cluster for the network at each time point (or each state), we want to reveal the trends in our data: we need to simplify and highlight the structural changes between clusters. In the mapping-change step of Fig. 2, we show how to construct an *alluvial diagram* of the example networks that highlights

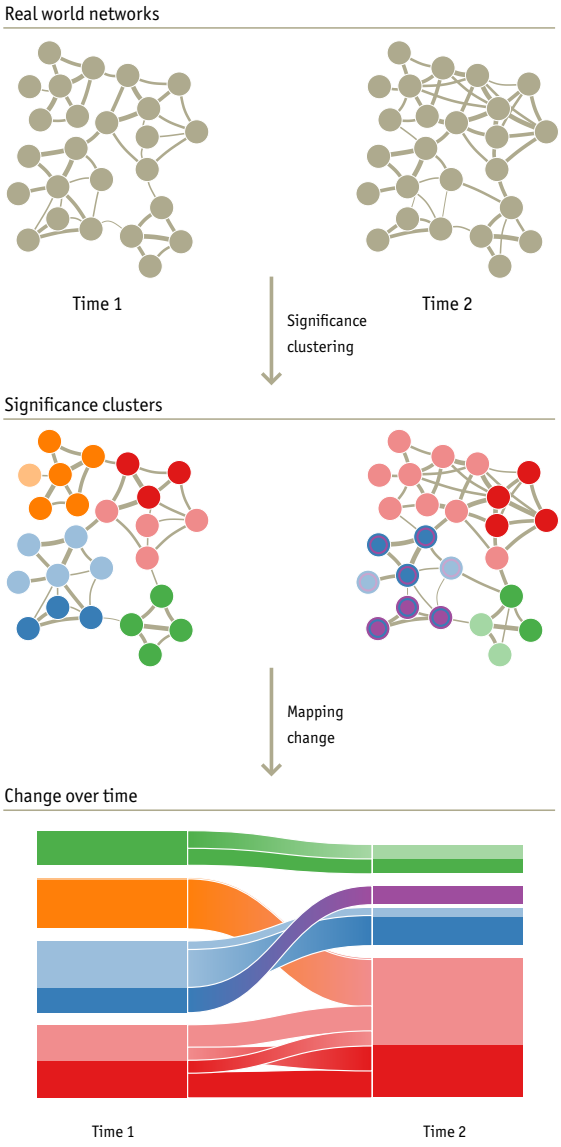

FIG. 2 Mapping change in networks. An alluvial diagram (bottom), with clusters ordered by size, reveals changes in network structures over time. Here the height of each block represents the volume of flow through the cluster (19), with significant subsets in darker color. The orange module merges with the red module, but the nodes are not clustered together in 95% of the bootstrap networks. The blue module splits, but the significant nodes in the blue and purple modules are clustered together in more than 5% of the bootstrap networks. With a 5% significance threshold, neither change is significant.

and summarizes the structural changes between the time 1 and time 2 significance clusters. Each cluster in the network is represented by an equivalently colored block in the alluvial diagram. Darker colors represent nodes that are assigned with statistical significance, while lighter colors represent insignificant assignments. Changes in the clustering structure from one time period to the next are represented by the mergers and divergences that occur in the ribbons linking the blocks at time 1 and time 2.

The alluvial diagram for the citation data reveals the significant structural changes that have occurred in science over the past decade. Rather than viewing the entire diagram, let us highlight a couple of interesting stories. Figure 3 shows a subset of biomedical fields for the years 2001, 2003, 2005, and 2007 (see the end of the supplement for all years and an additional alluvial diagram illustrating changes in the area of physics).

The alluvial diagram illustrates, for example, how over the years 2001–2005, urology gradually splits off from oncology and how the field of infectious diseases becomes a unique discipline, instead of a subset of medicine, in 2003. But these changes are just two of many over this period. In the same diagram, we also highlight the biggest structural change in scientific citation patterns over the past decade: the transformation of neuroscience from interdisciplinary specialty to a mature and stand-alone discipline, comparable to physics or chemistry, economics or law, molecular biology or medicine. In 2001, 102 neuroscience journals, lead by *the Journal of Neuroscience*, *Neuron*, and *Nature Neuroscience*, are assigned with statistical significance to the field of molecular and cell biology (dark orange, 84 of 102 journals are assigned significantly). Further, *Brain, Behavior, and Immunity*, *Journal of Geriatric Psychiatry and Neurology*, *Psychophysiology*, and 33 other journals appear with statistical insignificance in psychology (green, 6 of 36 journals are assigned significantly) and *Neurology*, *Annals of Neurology*, *Stroke* and 77 other journals appear with statistical significance in neurology (blue, 75 of 80 journals are assigned significantly). In 2003, many of these journals remain in molecular and cell biology, but their assignment to this field is no longer significant (light orange, 5 of 102 journals are assigned significantly). The transformation is underway. In 2005, neuroscience first emerges as an independent discipline (red). The journals from molecular biology split off completely from their former field and have merged with neurology and a subset of psychology into the significantly stand-alone field of neuroscience. (In 2006, shown in supplement, the structure reverts to a pattern similar to 2003.)

In their citation behavior, neuroscientists have finally cleaved from their traditional disciplines and united to form what is now the fifth largest field in the sciences (after molecular and cell biology, physics, chemistry, and medicine). Although this interdisciplinary integration has been ongoing since the 1950s (20), only in the last decade has this change come to dominate the citation structure of the field and overwhelm the intellectual ties along traditional departmental lines.

The problem of detecting structural change in large networks adds two new challenges to the basic method of network clustering: (1) we need appropriate statistical methods to identify significant features of network clustering and to distinguish between trends and noise in the data, and (2) we require effective visualizations to bring out the stories implicit in a time series of cluster maps. To resolve the first of these challenges, we have developed a

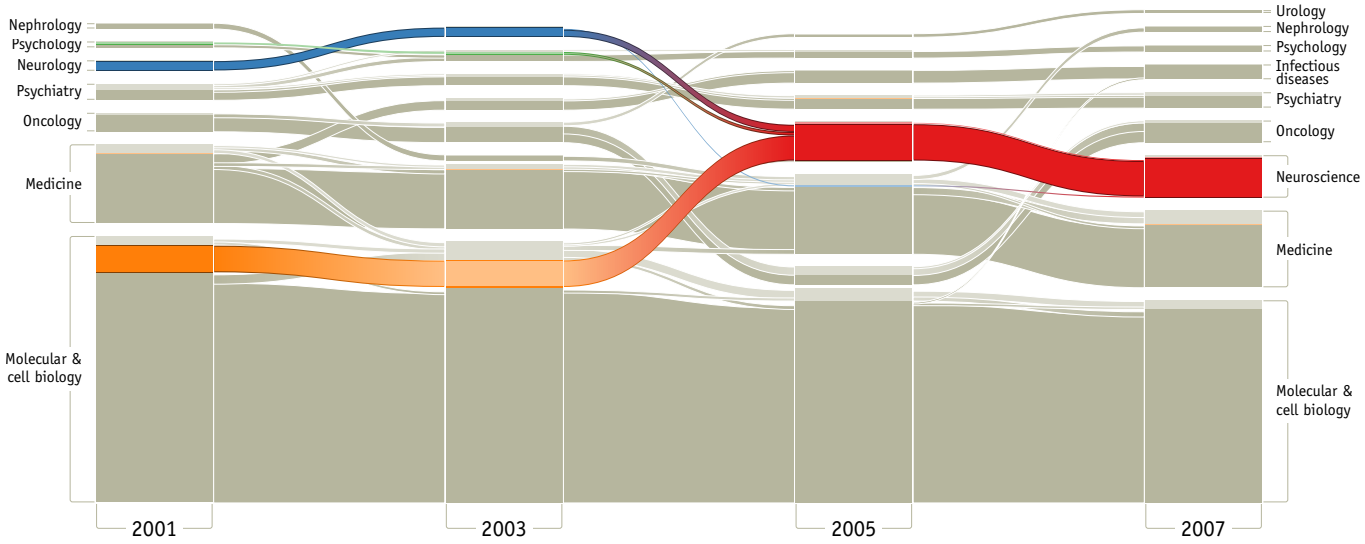

FIG. 3 Mapping change in science. This set of scientific fields show the major shifts in the last decade of science. Each significance clustering for the citation networks in years 2001, 2003, 2005, and 2007 occupies a column in the diagram and is horizontally connected to preceding and succeeding significance clusterings by stream fields. Each block in a column represents a field and the height of the block reflects citation flow through the field. The fields are ordered from bottom to top by their size with mutually nonsignificant fields placed together and separated by half the standard spacing. We use a darker color to indicate the significant subset of each cluster. All journals that are clustered in the field of neuroscience in year 2007 are colored to highlight the fusion and formation of neuroscience.

method for significance clustering based on the parametric bootstrap. To address the second, we have presented the visualization technique of alluvial diagrams. These methods are general to many types of networks and can answer questions about structural change in science, economics, and business.

Acknowledgments

The authors would like to thank Jevin West both for processing the journal citation data and for numerous helpful discussions, and Moritz Stefaner for his extensive help with the information visualizations presented here. This work was supported by the National Institute of General Medical Sciences Models of Infectious Disease Agent Study program cooperative agreement 5U01GM07649.

References

[1] M. Girvan, M. E. J. Newman, *Proc Natl Acad Sci USA* **99**, 7821 (2002).
[2] G. Palla, I. Derényi, I. Farkas, T. Vicsek, *Nature* **435**, 814 (2005).
[3] G. Palla, A. Barabasi, T. Vicsek, *Nature* **446**, 664 (2007).
[4] R. Guimerà, L. A. N. Amaral, *Nature* **433**, 895 (2005).
[5] M. Rosvall, C. T. Bergstrom, *PNAS* **105**, 1118 (2008).
[6] S. Fortunato, *arXiv:0906.0612* (2009).
[7] S. Thompson, *Sampling* (Wiley-Interscience, New York, 2002).
[8] R. A. Hanneman, M. Riddle, *Introduction to social network methods* (University of California, Riverside, CA, 2005).
[9] M. S. Handcock, D. R. Hunter, S. Goodreau, *J. Amer. Statistical Assoc.* **102**, 248 (2008).
[10] R. Albert, A. Barabási, *Rev Mod Phys* **74**, 47 (2002).
[11] M. E. J. Newman, *SIAM Review* **45**, 167 (2003).
[12] B. Efron, R. Tibshirani, *Monographs on statistics and applied probability* **57**, 1 (1993).
[13] D. J. de Solla Price, *Science* **149**, 510 (1965).
[14] Thomson-Reuters' Journal Citation Reports 1997-2007. Our data tally on a journal-by-journal basis the citations from articles published in a given year to articles published in the previous two years. Because we are interested in relationships between journals, we exclude journal self-citations.
[15] This *parametric* resampling of citations approximates a *non-parametric* resampling of articles, which makes no assumption about the underlying distribution. Currently we do not have access to article-level data.
[16] B. Karrer, E. Levina, M. E. J. Newman, *Phys Rev E* **77**, 046119 (2008).
[17] D. Gfeller, J.-C. Chappelier, P. D. L. Rios, *Phys Rev E* **72**, 056135 (2005).
[18] E. Costenbader, T. Valente, *Soc Networks* **25**, 283 (2003).
[19] This is the total *PageRank* of the cluster, which corresponds to the steady-state flow of random walkers that we use in the information-theoretic clustering algorithm.
[20] W. Cowan, D. Harter, E. Kandel, *Annu Rev Neurosci* **23**, 343 (2000).