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: December 8, 2008)

Change is the very nature of interaction patterns in biology, technology, economy, and science itself: The interactions within and between organisms change; the air, ground, and sea traffic change; the global financial flow changes; and the scientific research front changes. With increasingly available data, networks and clustering tools have become important tools to comprehend instances of these large-scale structures. But blind to the difference between noise and trends in the data, these tools alone must fail when used to study change. Only if we can assign significance to the partition of single networks can we distinguish structural changes from fluctuations and assess how much confidence should we have in the changes. Here we show that bootstrap resampling accompanied by significance clustering provides a solution to this problem. We use the significance clustering to realize de Solla Price’s vision of mapping the change in science.

Network analysis provides tools for understanding social and biological systems with numerous and diverse interacting components. For large networks, we need ways to highlight the important features while simplifying the overall structure. Researchers have developed a suite of network mapping tools for this purpose (1, 2, 3, 4); with them we can abstract, quantify, and comprehend the nature of a complex systems. Powerful as these tools have proven for understanding a system’s structure, we do not yet have an adequate tool for mapping how this structure changes. For example: 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? How has the network of global air traffic changed over the past half century? And how does science itself evolve as paradigms shift through time?

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, and the associated level of confidence, of large networks from observation of sample networks (5, 6, 7). But this comes at the cost of losing the unique identity of individuals. The reason that recent network approaches have become so prominent in the study of complex systems is that they capture and respect the identities and characteristics of the components (8, 9). Often these individual differences matter critically — and clustering rather than stratification must be deployed to comprehend the data (1, 2, 3, 4, 10).

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, in which Chicago O’Hare plays a unique and irreplaceable role. Because there is no way to look at multiple samples, the most effective approach to identify prominent and nonrandom features (11, 12), or to predict missing data (13, 14), is to compare the single networks to proper null models. While these approaches can tell us a lot about single networks, they do not allow us to map structural changes. To detect, highlight, and simplify significant structural changes over time or between states in large networks, we need to assess how much confidence we should have in clusterings of networks.

To assign significance to clusters of single networks, the bootstrap method is compelling (15). The bootstrap is a method for assessing the accuracy of an estimate by resampling from the empirical distribution of observations. But what do we do if we have only one observation, a single network? When the single observation is composed of numerous components, as is a network of nodes and links, the parametric bootstrap provides a solution. Instead of resampling directly from the empirical distribution, a parametric model is used to fit the data. For the networks of interest here, the identities of the nodes cannot be altered, parametrized, or resampled — it makes no sense to talk about the US air transit network without O’Hare, let alone with two O’Hares — but the link weights, which effectively define the nodes, can be parametrized and resampled without undermining the individual characteristics 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). Most importantly, we can reveal stories in network data. Here we illustrate by mapping...
change in the structure of science itself (16).

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, and through cumulative cycles of modeling and experimenting scientific research undergoes constant change: fields grow and shrink, merge and split. Citation patterns among scientific journals allow us to glimpse this flow of ideas and how the flow of ideas changes over time (16). Here we use journal aggregated citation data (18) from 1997 to 2006 and comprehend the networks with the information theoretic clustering method presented in ref. (4), which can reveal regularities of information flow across directed and weighted networks. We emphasize that, with the appropriate modification, the method of bootstrap resampling accompanied by significance clustering presented here is general and works for any type of network.

To assess the accuracy of a clustering, we resample a large number $B \sim 1000$ of bootstrap networks from the original network. For the directed and weighted citation

![Image](image_url)
network of science, we treat the citations as independent events and resample the weight of each link from a Poisson distribution with mean as in the original network (19) (see for example refs. (20, 21, 22) for other resampling techniques). Given the original network, Fig. 1 illustrates the clustering of this network and the clusterings of four of the bootstrap networks. When dealing with 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, but to assess the accuracy of the clusters requires a different approach.

To identify the nodes that are significant in their cluster assignments, we use simulated annealing to search for the largest significant subset of nodes 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 any other cluster, we search for all clusters whose significant subset are clustered with no other cluster’s significant subset in at least 95% of all bootstrap networks (see Method section). The first 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 need to simplify and highlight the structural changes between clusters and thereby bring out the stories in our data. In the second step of Fig. 2, we show how to construct an alluvial diagram that highlights and summarizes the structural changes between the time 1 and time 2 significance clusters. Each colored cluster in the network is represented by an equivalently colored block in the alluvial diagram. Solid colors represent significantly assigned nodes, while lighter colors represent insignificant assignments. Changes in the clustering structure from one time period to another are represented by the mergers and divergences that occur in the ribbons linking the blocks at time 1 and time 2.

To illustrate the power of this approach, we apply this method to citation data from Thomson-Reuters’ Journal Citation Reports. These data aggregate, at the journal level, approximately 60,000,000 citations among more than 7000 journals over the past decade. Comparing the significance clusters for each year by means of an alluvial diagram, we reveal the significant structural changes that have occurred in science over the past decade. Rather than viewing the entire diagram, let us pull out a couple of interesting stories. Fig. 3 show a subset of medical fields for the years 2001, 2003, and 2005.

As an illustrative example we describe the gradual integration of nephrology into medicine (purple stream in the diagram). In 2001, the field lead by Kidney International, Transplantation, and Journal of the American Society of Nephrology consists of 39 journals of which 33 are in the significant subset (89% of the citation flow) and the field as a whole is separated from all other fields in 99% of the clustered bootstrap networks. In the alluvial diagram, this is illustrated as a 89% dark purple block in the 2001 column. As illustrated by the unbroken stream of nephrology from 2001 to 2003, all nephrology journals in 2001 remain in the field in 2003, and no new additions join the field. Again in 2003 nephrology is clustered as a separate field and the significant subset now increases to 99% of the citation flow. But the field of nephrology is no longer significantly separated from the field of medicine. It is clustered together with medicine in 7% of the bootstrap networks, and therefore placed just under medicine in the alluvial diagram. In 2005 this fusion goes one step further and nephrology is no longer clustered as a separate field at all, but merges into medicine. This is illustrated by the stream that connects nephrology in 2003 to the nonsignificant subset of medicine in 2005.

The integration of nephrology is just one of many changes over this period. In the same diagram, we also highlight the biggest change in science 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 the majority of neuroscience journals (dark orange) are assigned with statistical significance to the field of molecular and cell biology. Others appear in psychology (green) and neurology (blue). In 2003, many of these journals (light orange) remain in molecular and cell biology, but their assignment to this field is no longer significant. The transformation is underway. In 2005, neuroscience first emerges 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 stand-alone field of neuroscience. (In 2006, not shown, the structure reverts to a pattern similar to 2003. It will be telling to observe what happens in 2007.) Neuroscience, which originated in the first studies of the nervous system more than a century ago and which for a long time existed as a set of independent disciplines, has now become a unified field. In their citation behavior, neuroscientists have cleaved from their traditional disciplines and united to form what is now the fifth largest field in the sciences of science (after molecular and cell biology, physics, chemistry, and medicine). Although this interdisciplinary integration has been ongoing since the 1950s (23), 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 in addition to the basic problem 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 method for significance clustering based on the parametric bootstrap. To address the second, we have presented the visualization technique of al-
FIG. 3 Mapping change in science. This set of scientific fields show the major shift in the last decade of science. Each significance clustering for the citation networks in years 2001, 2003, and 2005 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 by size. We use a darker color to indicate the significant subset of each cluster. The field of nephrology is highlighted to illustrate its merger with general medicine. All journals that a clustered in the field of neuroscience in year 2005 are colored to highlight the fusion and formation of neuroscience. All fields are equally spaced, except mutually nonsignificant fields, which are separated by half the standard spacing.

Alluvial diagrams. These method is general to many types of networks and can be applied to 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.

Method

Here we lay out the details of how we generate significance clusters and alluvial diagrams for mapping change in networks. Because this method assesses how much confidence we should have in a clustering of a network, we can detect, highlight, and simplify the significant structural changes that occur over time or between states in large networks, including but not limited to citation networks, traffic networks, and monetary flow networks. The method consists of four steps, summarized here and described in detail below:

1. We partition or “cluster” the original network, assigning each node to a single module or community of closely associated nodes.

2. We generate a large number (~ 1000) of bootstrap replicate networks, constructed by parametric bootstrap resampling of the original network. We cluster each of those networks.

3. To identify the significant assignments of nodes to modules in the original network, we search for the largest subset of nodes in the module that co-occur in at least 95 percent of all bootstrap networks. To identify the significant modules, we search for all modules whose significant nodes are clustered with no other module’s significant nodes in at least 95 percent of all bootstrap networks.

4. To map the changes in the network, we repeat the significance clustering for the different states of the network and generate an alluvial diagram, which
Significance clustering and alluvial diagrams

This approach to mapping change in large networks works for any clustering algorithm. The choice of algorithm depends on the network type (undirected, directed, unweighted, weighted) and the scope of the study. Here we focus on the general case of weighted directed networks. We also assume that the weight of the links can be described by a Poisson-like process. That is, the weights represent, or can be modeled by, independent events in time. This can be generalized to other distributions of link weights; see section 2 below.

For simplicity of description, here we map the change between two states $G^1$ and $G^2$ of a network — but it is straightforward to extend the procedure to more states. We enumerate the $N$ nodes by $\alpha = 1, 2, \ldots, N$. (The sets of nodes in $G^1$ need not be identical to the set in $G^2$.) By $w_{\alpha\beta}$ we denote a directed link from node $\alpha$ to node $\beta$ with weight $w$. Because the significance clustering procedure described below works exactly the same for each particular state of the network, in what follows we omit the superscript of $G$ unless necessary to avoid confusion.

1. Cluster real-world network

We first partition the network $G$ into the modular description $M$. In the modular description, each node is assigned to one and only one module. The number of modules depends on the network and the objective function of the clustering algorithm. To capture the dynamics across the links and nodes in directed weighted networks, we use the map equation as the objective function (4). In section Mapping directed weighted networks below, we present a new efficient algorithm to search for a partition of the network that minimizes the map equation. This search algorithm can also be generalized for other objective functions.

2. Generate and cluster bootstrap-world networks

The bootstrap is a statistical method for assessing the accuracy of an estimate by resampling from the empirical distribution. The method is particularly powerful when the variance of the estimator cannot be derived analytically or when the underlying distribution is not accessible. Because the cluster assignments are a result of a computational method and the network is idiosyncratic by nature, the bootstrap is indispensable for the process described here.

To generate a single bootstrap replicate network $G_b^*$, we resample every link weight $w_{\alpha\beta}$ of the original network $G$ from a Poisson distribution with mean equal to the original link weight $w_{\alpha\beta}$. That is, $w_{\alpha\beta}^* \sim \text{Pois}(w_{\alpha\beta})$ for each link in the bootstrap network. Because of the parametric resampling of the link weights, formally this method falls under parametric bootstrapping. If the link weights cannot be modeled by a Poisson process, or if the links are unweighted, the Poisson resampling should be replaced by an appropriate alternative resampling procedure (see for example refs. (20, 21)).

Subsequently we partition the bootstrap replicate network with the same clustering method as we used on the original network; this yields the bootstrap modular description $M_b^*$. This procedure — generating a bootstrap replicate network and clustering it into modules — is repeated to generate a large number $B \sim 1000$ of bootstrap modular descriptions $M^* = \{M_1^*, M_2^*, \ldots, M_B^*\}$. The panel Bootstrap world in Fig. 4 illustrates four of these modular descriptions for four different bootstrap replicate networks, each created by the Poisson resampling procedure described above. Because approximately 1000 networks must be clustered in this step, we have developed a new fast stochastic and recursive search algorithm for finding an accurate modular description of a given network (see section Mapping directed weighted networks).

3. Identify significant assignments

The basic idea behind significance clustering is that we can look at the bootstrap replicates to see which aspects of the modular description of the original network are best supported by the data. Features of the original network that occur in all or nearly all of the bootstrap replicates are well-supported; features that occur in only some of the bootstrap replicates are less well supported by the data.

What features do we consider? First, we consider the assignment of each node to a module. By looking at the set of bootstrap modular descriptions we can assess which of these assignments strongly supported by the data, and which node assignments are less certain. To identify the nodes that are significantly assigned to a module, we search for the largest subset of nodes in each module of the original modular description $M$ that are also clustered together in at least 95 percent of all bootstrap modular descriptions $M^*$. To pick the largest subset, we of course need some measure of size. The size of a subset could simply correspond to the number of nodes in the subset, but in line with our general clustering philosophy, we use the volume of flow through the subset.

To efficiently search the large space of possible subsets in each cluster, we use simulated annealing (24). Initially the nodes are randomly assigned to be members or

\footnote{This is the total PageRank of the subset, which corresponds to the steady-state flow of random walkers that we use in the information-theoretic clustering algorithm (4).}
FIG. 4  Significance clustering and alluvial diagram for mapping change in large networks.
non-members of the candidate largest subset. The score $S$ of the configuration is the size of the subset minus a penalty to account for the constraint that only nodes should be included that are clustered together in at least 95 percent of all bootstrap modular descriptions. To implement the penalty, we first, for each bootstrap modular description, count the number of nodes in the subset that do not belong to the largest group of nodes assigned to the same cluster. These are the mismatch nodes that break the constraint. To allow for a five percent error, we add together the number of mismatch nodes for all bootstrap modular descriptions, excepting the five percent with the highest number of mismatches. Finally we multiply this sum by ten times the cluster size. This we do to make sure that the subset size and the penalty are of comparable size, which is necessary for an efficient search and a zero penalty in the end of the procedure (this ad hoc scaling factor of 10 was found by optimizing the convergence to a configuration with zero penalty and maximal subset size). After initiating with random assignments, we follow the standard simulated annealing scheme (24). At successively lower temperatures $T$, a node’s subset assignment (member or non-member) is flipped and the score $S’$ for the new state is calculated. As in the Metropolis-Hastings algorithm (25, 26), the new state is always accepted if the new score is higher ($ΔS = S’ - S > 0$) or, if the new score is lower, the new state is accepted with probability equal to the Boltzmann factor of the score difference $\exp(ΔS/T)$. Starting at $T = 1$, we iterate this step as many times as there is nodes in the cluster, and then reduce the temperature according to $T’ = 0.99T$. We repeat this procedure for as long as at least one new state is accepted for a given temperature. The nodes assigned to the subset at the final state serves as our approximation for the largest significant subset.

In additional to telling us about the assignment of individual nodes to specific modules, the set of bootstrap replicates also contains information about which modules stand alone and which are possibly subsets of other modules. To reveal this information we need to identify the modules that are always, or almost always, separate from any other module. We consider a module to be significant if its significant subset is clustered with no other significant subset in at least 95 percent of all bootstrap modular descriptions. Conversely, two clusters are mutually nonsignificant if their significant subsets are clustered together in more than 5 percent of all bootstrap modular descriptions. In this way, each module can be mutually nonsignificant with more than one other module. In the alluvial diagram described below, we want to associate each nonsignificant module with the module together with which it most likely form a subset. The search for these pairs of modules is straight forward: For each pair of modules, we count in how many bootstrap modular descriptions all nodes in the two significant subsets are clustered together and record this number if it exceeds 5 percent of all bootstrap modular descriptions (the criterion for nonsignificant modules). Then, starting at the smallest module, we associate the module with the other larger module that it is most often clustered with, and proceed to the next smallest module and so on.

4. Construct alluvial diagram

To reveal change over time or between states of real-world networks, we summarize the results of the significance clusterings of the different states $G^1, G^2, \ldots$ in an alluvial diagram. The diagram is constructed to highlight the significant changes, fusions, and fissions that the modules undergo between each pair of successive states $G^i$ and $G^{i+1}$. For this reason, each significance clustering for a state $G^i$ occupies a column in the diagram and is horizontally connected to preceding and succeeding significance clusterings by stream fields. Each block in a row of the alluvial diagram represents a cluster and the height of the block reflects the size of the cluster (here in units of flow through the cluster, though other size measures, such as number of nodes, could be used instead). The modules are ordered by size, or if higher-order module structure exists, they are ordered by size within each super-module. We use a darker color to indicate the significant subset of each cluster. Different colors can be used for clusters or groups of clusters to highlight particular stories in the data. All clusters are equally spaced, except mutually nonsignificant clusters, which are separated by half the standard spacing.

We use the stream fields to reveal the changes in cluster assignments and in level of significance between two adjacent significance clusterings. The height of a stream field at each end, going from the significant or nonsignificant subset of a cluster in one column to the significant or nonsignificant subset of a cluster in the adjacent column, represents the total size of the nodes that make this particular transition. By following all stream fields from a cluster to an adjacent column, it is therefore possible to study in detail the mergers with other clusters and the significance transitions. To reduce the number of crossing stream fields, the stream fields are ordered by the order of the connecting clusters.

To keep related stream lines together, we let them pass through the mid point between the mid points of the exiting and entering subsets. For smooth transitions, we draw the stream fields with splines and use gradient shading for the component colors. Finally, to reduce the amount of ink and improve clarity, the stream lines have a slim waist.

Mapping directed weighted networks

Here we briefly review our information theoretic approach to reveal community structure in weighted and directed networks (4) and present a new fast stochastic and recursive search algorithm to minimize the map equa-
tion, the objective function of our method. This method we have developed to be able to accurately partition the large number of bootstrap networks. The search algorithm can also be generalized for other objective functions.

The map equation

The objective of our information theoretic method is to partition the nodes of a network into modules so as to minimize the expected description length of a random walk across the nodes and links of the network. For a given partition, the expected description length is quantified by the map equation. For a detailed description of the map equation and this method, see the supporting appendix of ref. (4). Here follows a short review.

Define a module partition $M$ as a hard partition of a set of $n$ nodes into $m$ modules such that each node is assigned to one and only one module. The map equation $L(M)$ gives the average number of bits per step that it takes to describe an infinite random walk on a network partitioned according to $M$:

$$L(M) = q_\infty H(Q) + \sum_{i=1}^{m} p_i H(P_i).$$

The map equation calculates the minimum description length of a random walk on the network for a two-level code that separates the important structures from the insignificant details based on the partition $M$. This two-level code uses unique codewords to name the modules specified by partition $M$, but reuses the codewords used to name the individual nodes within each module. The first term of this equation gives the average number of bits necessary to describe movement between modules, and the second term gives the average number of bits necessary to describe movement within modules. In the first term, $q_\infty$ is the probability that the random walk switches modules on any given step and $H(Q)$ is the entropy of the module names. In the second term, $H(P_i)$ is the entropy of the within-module movements — including an “exit code” to signify departure from module $i$ — and the weight $p_i$ is the fraction of within module movements that occur in module $i$, plus the probability of exiting module $i$ such that $\sum_{i=1}^{m} p_i = 1 + q_\infty$.

To efficiently describe a random walk using a two-level code of this sort, the choice of partition $M$ must reflect the patterns of flow within the network, with each module corresponding to a cluster of nodes in which a random walker spends a long period of time before departing for another module. To find the best such partition, we therefore seek to minimize the map equation over all possible partitions $M$.

### Fast stochastic and recursive search algorithm

Any greedy (fast but inaccurate) or Monte Carlo based (accurate but slow) approach can be used to minimize the map equation. But since on the order of 1000 networks must be clustered for each significance clustering and high accuracy always is desirable, we have developed a new method that provides a good balance between the two extremes. As a reference, the new algorithm is in practice as fast the our previous high-speed algorithms (the greedy search presented in the the supporting appendix of ref. (4)), which was based on the method introduced in ref. (27) and refined in ref. (28). Yet it is also more accurate than our previous high-accuracy algorithm (a simulated annealing approach) presented in the same supporting appendix.

The core of the algorithm follows closely to the method presented in ref. (29): neighboring nodes are joined into modules, which subsequently are joined into super modules and so on. First, each node is assigned to its own module. Then, in random sequential order, each node is moved to the neighboring module that results in the largest decrease of the map equation. If no move results in a decrease of the map equation, the node stays in its original module. This procedure is repeated, each time in a new random sequential order, until no move generates a decrease of the map equation. Now the network is rebuilt, with the modules of the last level forming the nodes at this level. And exactly as at the previous level, the nodes are joined into modules. This hierarchical rebuilding of the network is repeated until the map equation cannot be reduced further. Except for the random sequence order, this is the algorithm described in ref. (29).

With this algorithm, a fairly good clustering of the network can be found in a very short time. Let us call this the core algorithm and see how it can be improved. The often large number of nodes assigned to the same module are forced to move together once the network is rebuilt and what was an optimal move early in the algorithm might have opposite effect later in the algorithm. Because two or more modules that merge together and form one single module when the network is rebuilt can never be separated again in this algorithm, the accuracy can be improved by extending the core algorithm by breaking the modules of the final state in any of the two following ways:

**Submodule movements.** First each cluster is treated as a network on its own and the main algorithm is applied to this network. This procedure generates for each module one or more submodules. Then all submodules are moved back to their respective modules of the previous step. At this state, with the same partition as in the previous step but with each submodule being freely movable between the modules, the main algorithm is re-applied.

**Single-node movements.** First each node is first reassigned to be the sole member of its own mod-
ule, in order to allow for single-node movements. Then all nodes are moved back to their respective modules of the previous step. At this state, with the same partition as in the previous step but with each single node being freely movable between the modules, the main algorithm is re-applied.

In practice we repeat the two extensions to the core algorithm in sequence and repeatedly as long as the clustering is improved. Moreover, we apply the submodule movements recursively. That is, to find the submodules to be moved, the algorithm first splits the submodules into subsubmodules, subsubsubmodules, and so on until no further splits are possible. Finally, because the algorithm is stochastic and fast also with the two extensions, we can restart the algorithm from scratch every time the clustering cannot be improved further and the algorithm stops. The implementation is straightforward and makes the final partition less likely to come from a local minimum. For each iteration, we record the clustering if the description length is shorter than the previously shortest description length. In practice, for the citation networks presented in this paper, which have on the order of 10,000 nodes and 1,000,000 directed and weighted links, each iteration takes around 5 seconds on a modern PC. We generate the significance clusterings by repeating the algorithm 100 times for each network and bootstrap network.

References

5. R. A. Hanneman, M. Riddle, Introduction to social network methods (University of California, Riverside, CA, 2005).
17. 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.
18. Journal Citation Reports 1997-2006, Thomson Scientific. Our data tally on 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.
19. 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.