Tracing the evolution of physics on the backbone of citation networks

S. Gualdi,[1] C. H. Yeung,[1,2] and Y.-C. Zhang[1,3]

[1]*Department of Physics, University of Fribourg, CH-1700 Fribourg, Switzerland*

[2]*The Nonlinearity and Complexity Research Group, Aston University, Birmingham B4 7ET, United Kingdom*

[3]*Web Sciences Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China*

(Received 5 August 2011; published 14 October 2011)

Many innovations are inspired by past ideas in a nontrivial way. Tracing these origins and identifying scientific branches is crucial for research inspirations. In this paper, we use citation relations to identify the *descendant chart*, i.e., the family tree of research papers. Unlike other spanning trees that focus on cost or distance minimization, we make use of the nature of citations and identify the most important parent for each publication, leading to a treelike backbone of the citation network. Measures are introduced to validate the backbone as the descendant chart. We show that citation backbones can well characterize the hierarchical and fractal structure of scientific development, and lead to an accurate classification of fields and subfields.

DOI: 10.1103/PhysRevE.84.046104 PACS number(s): 89.75.Hc, 89.75.Fb, 02.50.−r, 05.45.Df

I. INTRODUCTION

Many innovations are inspired by past ideas in a nontrivial way. Examples in statistical physics include the connection between spin glasses and combinatorial problems [1,2], the application of critical phenomena in earthquake modeling [3,4], and the analyses of the spread of disease using percolation theory [5,6]. To draw these connections is easy, but to map individual fields onto a *descendant chart*, i.e., a family tree of research branches, is more complicated. An even more difficult task is to uncover the macroscopic tree based on the microscopic relations between publications. Despite the difficulties, descendant charts are crucial for revealing the nontrivial connections between branches, which stimulate inspirations. Accurate descendant charts also provide a natural classification of papers.

A solid basis to study descendant charts is represented by the citation network, which can be seen as the original map of scientific development. In recent years, citation and authorship networks have been used to evaluate the impact of academic papers and scientists [7,8]. Although useful information is retrieved, most studies focus on contemporary impact and ignore the intrinsic hierarchical structure of citations encoding the generation of scientific advances. Unlike the horizontal exploration in conventional paper classifications [9], we explore the vertical, i.e., temporal, dimension in citation networks to identify the descendant charts of publications.

Toward this end, we identify a *backbone* of the citation network by removing all but the most relevant citation for each paper. The backbone hence results in a treelike structure and is found solely based on citation relations with no additional information. Similar concepts of spanning trees are extensively studied in transportation networks and oscillator networks, as minimum spanning trees in terms of traveling cost [10,11], and trees that maximize betweenness [12] or synchronizability [13]. Although the citation backbone can be constructed by these definitions, we see no direct correspondence between them and scientific descendant trees. Instead, one should make use of the nature of citation relations and identify the important parent and thus the offspring for each paper, which constitutes a backbone specific for citation networks.

In this paper, we identify the descendant chart for publications in journals of the American Physical Society (APS), based on their citation network from 1893 to 2009. Our objectives are threefold. First, we introduce a potential approach to identify the most relevant parent (among the set of original references) for each publication, which leads to a backbone of the citation network. Second, we introduce measures to validate the citation backbones as representative descendant charts and compare our approach with two other simple procedures (i.e., the selection of a random parent or the longest path to the root). Finally, we show that citation backbones possess features of hierarchy and self-similarity, and lead to a valid classification of papers in linear time, compared to conventional polynomial-time algorithms [14,15]. The present work pinpoints the importance of scientific descendant charts, as well as their intrinsic difference from other spanning trees.

II. METHODS

To start our analyses, we first denote the references of a paper as its *parents*, and the articles citing the paper as its *offspring*. The set of parents and the offspring of a paper i are denoted by $\mathcal{P}_i$ and $\mathcal{O}_i$ with p_i and o_i elements, respectively. Intuitively, the offspring of an important paper should share a similar focus introduced by its influential parent. Less relevant parents will, by contrast, lead to a more heterogeneous descendance. We thus quantify the *impact* of parent α on i by $I_{\alpha\to i} = \sum_{i'\in\mathcal{O}_\alpha\setminus\{i\}} s_{ii'}$, where $s_{ii'}$ is the *similarity* between i and i'. We refer to papers in the set of $\mathcal{O}_\alpha\setminus\{i\}$ as the *peers* of i rooted in α (see Fig. 1 for an illustration). The higher the overall similarity between i and the papers in $\mathcal{O}_\alpha\setminus\{i\}$, the higher the impact of α on i.

A simple way to measure the similarity between i and peer i' is to count the number of their common references, i.e., $s_{ii'} = |\mathcal{P}_i \cap \mathcal{P}_{i'}|$. However, this similarity measure favors peers with many references, resulting in an impact biased toward parents with a large offspring. This suggests to define a similarity measure based on a random walk from the peers to i. We thus consider a two-step random walk from each peer i' to i

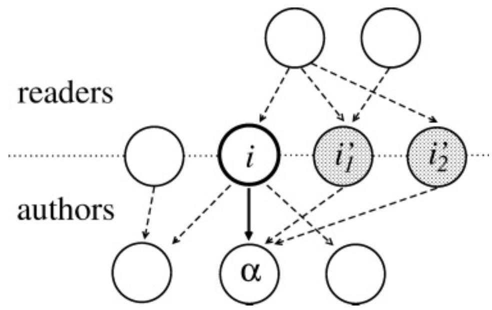

FIG. 1. A schematic diagram that shows two peers $i' = i'_1, i'_2$ (shaded) of i rooted from parent α. To compute each $s^{\text{aut}}_{ii'}$, we consider a random walk from i' through papers cited by both i' and i. Specularly, to compute each $s^{\text{read}}_{ii'}$ we consider a random walk from i' through papers citing both i' and i.

that passes through their common references, and we define a contribution to $s_{ii'}$ as

$$s^{\text{aut}}_{ii'} = \frac{1}{p_{i'}} \sum_{j \in \mathcal{P}_i \cap \mathcal{P}_{i'}} \frac{1}{o_j}. \tag{1}$$

The superscript represents the *authors*' interpretations, as this similarity is measured through the references chosen by the authors of i. A second contribution is instead given by a random walk through articles citing i and represents the *readers*' interpretation of i. In analogy with $s^{\text{aut}}_{ii'}$, we define

$$s^{\text{read}}_{ii'} = \frac{1}{o_{i'}} \sum_{j \in \mathcal{O}_i \cap \mathcal{O}_{i'}} \frac{1}{p_j}. \tag{2}$$

As defined in both Eqs. (1) and (2), the higher the random-walk probability from i' to i, the higher the similarity between i' and i.

Finally, by combining linearly $s^{\text{read}}_{ii'}$ and $s^{\text{aut}}_{ii'}$ and summing over all the peers rooted in α, we obtain the impact of α on i as

$$I_{\alpha \to i} = \sum_{i' \in \mathcal{O}_\alpha \setminus \{i\}} \left[f s^{\text{read}}_{ii'} + (1-f) s^{\text{aut}}_{ii'} \right] \tag{3}$$

with f to adjust the relative weights on the two contributions. The subsequent analysis is simplified by setting $f = 0.5$ unless otherwise specified. We note that citations between peers [16] do not contribute to the above measure, as these citations may correspond to relations other than similarity. For instance, if many peers rooted from parent α cited i, it implies that α complements i instead of being merely an influential parent of i. The same is true if i cites many peers rooted from α, which suggests α is a complement of its peers instead of a mere influential parent.

By keeping only the reference α with highest $I_{\alpha \to i}$ for all i, we obtain a citation backbone denoted as the *SIM* backbone. Cases of equal scores are extremely rare and do not affect results (in such situations, we arbitrarily choose the latest reference with the highest $I_{\alpha \to i}$). In addition, we examine also the *RAN* and the *LON* backbone, which selects, respectively, a random parent and the reference giving rise to the longest path to the root (most likely the latest published parent). Other than serving as a benchmark, the *RAN* backbone can be informative as the random parent is one of the original references. The *LON* backbone instead represents a natural choice if progress is always based on recent developments, as one may follow the step-by-step evolution of science represented by the maximum number of steps needed to reach the root.

III. STATISTICAL PROPERTIES OF THE BACKBONE

We will examine the citation network among the journals of the American Physical Society, from the years 1893 to 2009. The dataset is composed of 4.67×10^6 citations between 4.49×10^5 publications. In rare cases there are references to contemporary or even posterior published papers. These citations are removed and the network is strictly acyclic.

We note that all papers without reference are potentially the roots of the backbone. As this number is generally greater than 1 and we are limited to the simplest case with one selected ancestor per paper, there may appear multiple roots and hence isolated trees in the backbone. In the subsequent discussion, we will refer to the output of the *SIM*, *RAN*, and *LON* algorithms as *backbone*, and its isolated components as *trees*. Since the seemingly isolated roots may be connected by citations other than the APS network, the number of isolated trees would be lower if more comprehensive citation data were used. Nevertheless, such isolated trees may represent a crude classification of papers. Table I summarizes general statistical properties of the group of trees as obtained by *SIM*, *RAN*, and *LON* approaches.

IV. THE STRUCTURE OF THE BACKBONE

In this section, we will discuss and derive measures to validate the citation backbones as representative of descendant charts. Three aspects will be studied. First, we examine the linkage between different generations of papers. Second, we quantify the paper classification as given by the clustering and branching structure in the backbones. Finally, we examine the possible self-similarity in citation backbones.

A. Hierarchy

We first examine the probability of observing an original citation between two papers as a function of their distance in the backbone. If the backbone is meaningful, we expect this quantity to decrease fast as the distance increases. To compute the distance between i and j, we find the first common ancestor

TABLE I. Statistical properties of the isolated trees in the *SIM*, *RAN*, and *LON* backbones. Values for *RAN* are averaged over 100 realizations. We also show the average interval (in years) between the date of publication of a paper and its parent.

	SIM	*RAN*	*LON*
Number of isolated trees	3953	6594	2630
Size:			
largest tree	26 115	30 358	428 147
2nd largest tree	25 386	15 697	592
3rd largest tree	21 794	11 362	471
$\langle \Delta t \rangle$ parent-offspring	9.5 y	7.4 y	2.1 y

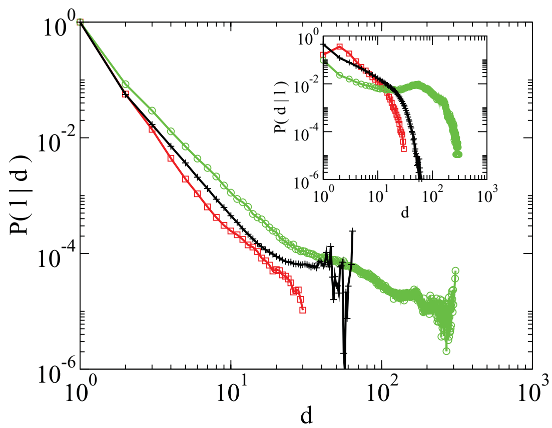

FIG. 2. (Color online) Conditional probability $P(l|d)$ of observing a citation between two papers at distance d on the *SIM* (red squares), the *RAN* (black plus), and the *LON* (green circles) backbone. Results for *RAN* are averaged over 100 realizations of the backbone (the variance is negligible). Inset: conditional probability $P(d|l)$ that two papers are at distance d assuming there is a citation.

α' in the backbone and count the number of steps $d_{i\alpha'}$ and $d_{j\alpha'}$ required to go from i to α' and from j to α'. The distance d_{ij} is then set as $d_{ij} = d_{i\alpha'} + d_{j\alpha'}$. We consider $d_{ij} = \infty$ for paper i and j in isolated trees. In Fig. 2, we plot $P(l|d)$ as a function of d for all *SIM*, *RAN*, and *LON* backbones, where l denotes the presence of a link, i.e., a citation. As we can see, $P(l|1) = 1$ by definition and all $P(l|d)$ display a power-law decay for small d. The *SIM* backbone shows a faster decay than other algorithms, suggesting that citations are more localized in the neighborhood of a paper in the *SIM* backbone. A similar quantity $P(d|l)$ (see the inset of Fig. 2) also indicates that the *SIM* backbone is the best representative of the APS network since citations are concentrated at $d = 2$ and decay faster as the distance increases.

In addition to $P(l|d)$, we further consider $P(l|d_{i\alpha'}, d_{j\alpha'})$, where α' is again the first common ancestor of i and j in the backbone. This allows us to see whether citations are localized on the specific branch of each paper or spread over different ramifications on the tree. For any pair (i,j), we take i as the later published paper such that the only potential citation is $i \to j$. We show in Figs. 3(a)–3(c) the results of $P(l|d_{i\alpha'}, d_{j\alpha'})$ for the three backbones as a function of $d_{i\alpha'}$ and $d_{j\alpha'}$. One notes that increasing $d_{i\alpha'}$ on the line of $d_{j\alpha'} = 0$ corresponds to the vertical trace toward the root, while points with $d_{j\alpha'} \neq 0$ correspond to the various "ramifications" in the backbone. Both *SIM* and *RAN* give a meaningful structure where citations are localized on the descendant chart of the immediate and next immediate ancestor, i.e., the triangle in the bottom left-hand corner. Citations between different ramifications are rare. The *LON* backbone instead displays a less coherent structure in which citations crossing different lines of research are common. To examine the difference between *SIM* and *RAN*, we also show the scaled difference of their $P(l|d_{i\alpha'}, d_{j\alpha'})$ as given in the vertical axis of Fig. 3(d). This comparison clearly indicates that *SIM* gives rise to the

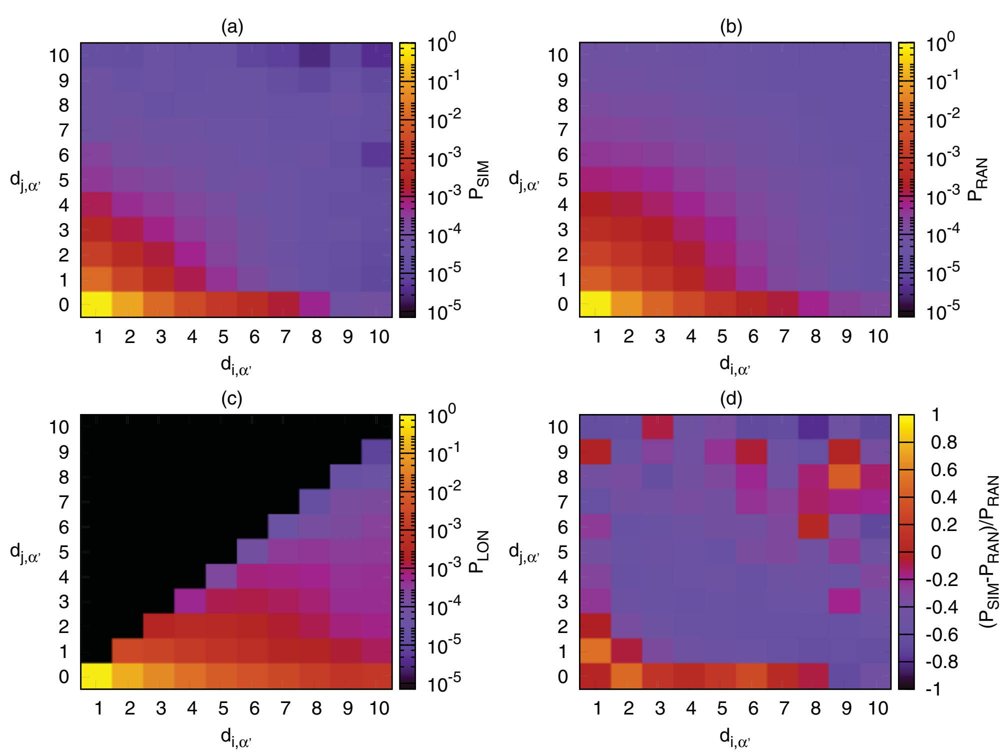

FIG. 3. (Color online) Heat maps that show $P(l|d_{i\alpha'}, d_{j\alpha'})$ as a function of $d_{i\alpha'}$ and $d_{j\alpha'}$ for (a) the *SIM* backbone, (b) the *RAN* backbone, and (c) the *LON* backbone, with citation $i \to j$. Since papers only cite references published before them, the observed dark triangle in *LON* suggests a rather homogeneous temporal interval between papers and their best *LON* ancestor, such that citations with $d_{j\alpha'} > d_{i\alpha'}$ are highly improbable. Results for *RAN* are averaged over 100 realizations of the backbone (the variance is negligible). (d) Scaled difference of $P(l|d_{i\alpha'}, d_{j\alpha'})$ between *SIM* and *RAN* .

most meaningful hierarchy as citations are mainly found on the descendant chart of the more relevant ancestors instead of crossing different charts.

B. Clustering

In addition to the crude classification as given by the isolated trees, the branches in a single tree are also informative to identify research fields and subfields. From the clustering point of view, the method we have introduced is computationally efficient [with complexity $O(N)$ as long as connectivity is not extensive] compared to modularity maximization-based algorithms [17,18] or hierarchical clustering algorithms [19] [with complexity at least $O(N^2)$]. Moreover, the clustering naturally explores the temporal dimension by preserving the ancestor-descendant relations.

In order to map the backbone into clusters, we consider two simple approaches that involve only a single parameter. The first approach makes use of the publication year of papers and naturally follows the order of publication. We first make a cut at the year Y_c such that papers printed before Y_c are removed. We then consider each unconnected component as a different branch, i.e., a different cluster, in the original backbone, and as a classification for papers.

The second approach is dependent on the cluster size, which we consider to be a typical research branch. Starting from the leaves of the backbone (i.e., papers with no offspring), we trace toward the root until a branching point is reached. The branching point is defined as a node of the network from which at least (i) two ramifications start and (ii) two ramifications are extended more than S steps. When a branching point satisfies these requirements, all ramifications originating from it are considered as different branches, resulting in a classification of papers. Here we quantify the validity of clustering as a function of parameter Y_c and S.

In order to evaluate the quality of a given clustering, we use two different measures. The first one—which we call *exclusivity*—is a modified modularity measure specific for directed acyclic graphs. The rationale behind this measure is to compute the fraction of links of the original network falling inside the same cluster and compare it with the expected value for a random directed acyclic graph. We denote the set of papers assigned to branch x as $\mathcal{X}$ and define the *exclusivity* as

$$E = \left\langle \left\langle \frac{p_x^i}{p_i} - \frac{n_x^i}{n_i} \right\rangle_{i\in\mathcal{X}} \right\rangle_x, \tag{4}$$

where p_i is again the number of references of i, p_x^i is the number of i's references in branch x, n_i is the number of papers published before i, and n_x^i is the number of papers published before i in branch x. The term n_x^i/n_i thus corresponds to the expected fraction of links from i to an element in $\mathcal{X}$ in the random case. To reduce the noise from small clusters, we have excluded branches with fewer than 10 papers.

The second measure we use is the effective number of PACS—N_P—which counts the average number of heterogeneous PACS in individual branches. Good paper classifications result in small values of N_P. We first denote r_p^x to be the fraction of paper in branch x, which is labeled by the PACS number p, and note that $\sum_p r_p^x \geqslant 1$ as papers are always labeled by more than one PACS number. N_P is then defined as

$$N_P = \left\langle \frac{1}{\sum_p \left(f_p^x\right)^2} \right\rangle_x, \tag{5}$$

where $f_p^x = r_p^x / \sum_{p'} r_{p'}^x$. Therefore, $N_P = 1$ when there is only one PACS in the branch, which corresponds to the optimal classification of papers. On the other hand, N_P attains its maximum when all PACS numbers in $\mathcal{X}$ have an equal share (i.e., equal f_p^x) and a large N_P thus corresponds to high heterogeneity inside single clusters. We remark that in evaluating N_P, only the first four digits are used to distinguish PACS numbers.

In Fig. 4, we plot the E and N_p as a function of the two parameters Y_c and S. Both measures are biased by the cluster size but in an opposite way. While N_p indicates better clustering (and thus a lower value) when isolated clusters are of a smaller size, E indicates better clustering (and thus a higher value) when clusters are of a larger size. Even with the compensation by n_i^x/n_i in Eq. (5), we still observe a small bias of E on cluster size. These biases may influence our comparison of the identified clusters from the *SIM*, *RAN*, and *LON* backbones, as they have different sizes. Nevertheless, the combination of the two independent measures clearly indicates that *SIM* is the best choice to obtain a meaningful clustering besides the bias introduced by cluster sizes. Moreover, the exclusivity of the *SIM* backbone is higher for any value of the parameter S, which further supports the validity of the comparison despite the presence of the bias.

C. Self-similarity

Other than the hierarchical and clustering properties, the backbones may possess self-similarity. Intuitively, self-similarity may be induced when branches of research successively generate branches of significant advances. The existence of fractality in the backbone would provide support for its relevance with the evolution of science.

To show the self-similarity in networks, one can measure their fractal dimension by the box-covering method [12,20,21]. In this approach, the fractal dimension d is defined as the power-law exponent in

$$N(l_B) \sim l_B^d, \tag{6}$$

where $N(l_B)$ is the minimum number of boxes, each of radius l_B, required to cover the whole network. To obtain the exact $N(l_B)$ is difficult, thus we employ the random sequential box-covering algorithm [21], which gives an approximate $N(l_B)$ with the same scaling. Specifically, we start with all nodes being "uncovered" and repeat the following procedures until all nodes become "covered": (i) randomly pick a seed node, (ii) find all "uncovered" nodes within a distance of l_B from the seed, and (iii) increase $N(l_B)$ by 1 if there exists at least one "uncovered" node and mark all of them as "covered." Note that a "covered" node can also be a seed in the subsequent searches. For the same tree, we show the minimum of $N(l_B)$ among 20 random sequences as our final value for each value of l_B.

We show in Fig. 5 the results of $N(l_B)$ as a function of l_B for the largest tree in the *SIM*, *RAN*, and *LON* backbone. The

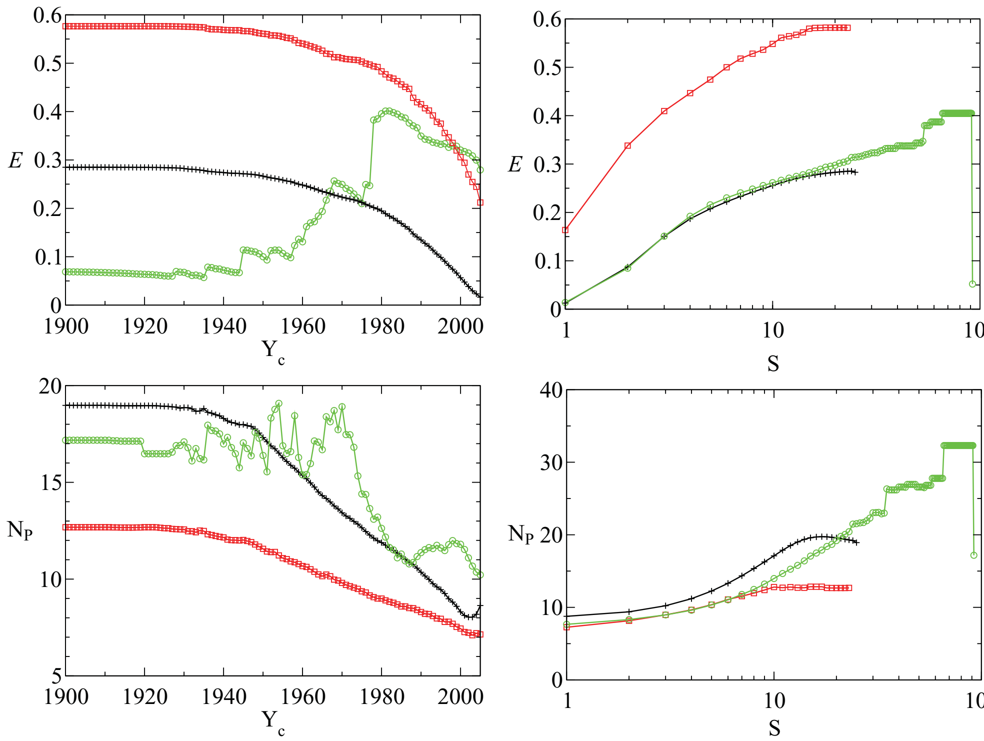

FIG. 4. (Color online) The exclusivity E and the effective number of PACS N_P as a function of cut-year Y_c and branch depth S for the *SIM* (red squares), the *RAN* (black plus), and the *LON* (green circles) backbone. Both quantities show that *SIM* gives a more meaningful division into branches. Results for *RAN* are averaged over 100 realizations of the backbone (the variance is negligible).

results are compared to $N(l_B)$ of the original citation network. As we can see, $N(l_B)$ from the *LON* backbone has the greatest resemblance to power laws, while that of the *RAN* backbone shows the fastest decay in $N(l_B)$. The *LON* tree has a long tail of $N(l_B)$, as it is longest and largest in size (see Table I). Only the largest tree of a particular realization of the *RAN* backbone is shown, as similar results are observed in other realizations. Although a long tail is not observed in the *SIM* tree, it shows a power-law-like behavior up to an intermediate value of l_B.

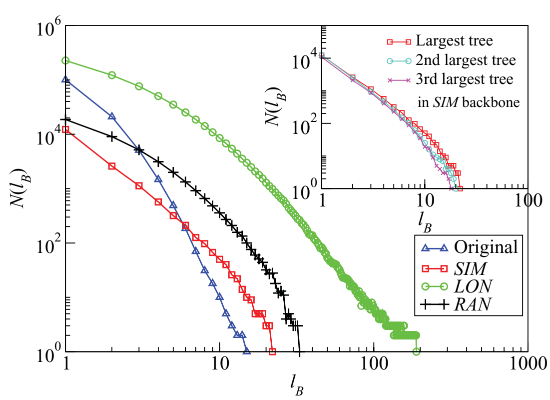

FIG. 5. (Color online) $N(l_B)$ as a function of l_B for the largest tree in the original network and its *SIM*, *LON*, and *RAN* backbone, taken as the minimum over 20 random sequences of seed nodes for the box-covering algorithms. The *SIM* backbone is obtained at $f = 0.5$. Inset: $N(l_B)$ as a function of l_B for the three largest trees in the *SIM* backbone.

Similar behaviors are also observed in the other isolated trees of the *SIM* backbone, as shown in the inset of Fig. 5.

We interpret the results as follows. The observed resemblance to power laws from the *SIM* and *LON* backbone may suggest the presence of self-similarity in their descendant chart. While the *LON* backbone does not possess a meaningful hierarchy or clustering compared to the *SIM* backbone, its step-by-step structure indeed shows the highest fractality. We note that a rather short power law is also observed in the original network, though characterized by a different exponent from the *SIM* and *LON* backbones. On the other hand, such fractality is not observed in the *RAN* backbone.

V. POTENTIAL APPLICATIONS

In this section, we briefly describe the implications and potential applications of the citation backbone as a descendant chart of research papers.

As the backbone is a sketch of the skeleton of scientific development, it can be applied to identify seminal papers. Preliminary results show that a simple measure based on the the number of *relevant offspring*, i.e., followers in the backbone, is sufficient to give a meaningful ranking that is not trivially correlated with the original number of incoming citations (between the two rankings, Kendall's correlation coefficient is 0.19 and there is an overlap of only seven papers in the top 20 ranks). This serves as a simple yet meaningful definition of the impact of a publication. More refined definitions that take into account the reputation of each relevant offspring and/or the structural role of a given paper in the backbone can give an even better selection of fundamental papers. Moreover, our formulation of tunable weight on authors' and readers' interpretations in Eq. (3) can be easily incorporated in common ranking algorithms such as Page Rank, where an even repartition of citation importance is assumed instead.

The second application corresponds to the classification of papers. As we have mentioned before, such clustering divides papers into research fields or subfields and offers a basis for a synthetic picture of the state-of-the-art. There are several advantages over conventional classifications, including (i) lower computational complexity, (ii) additional information of subclustering as given by the internal tree structure, and (iii) predictions of future development by considering the rate of growth of subbranches. This last feature is particularly useful to filter the most active directions in the large amount of literature at our disposal.

VI. CONCLUSIONS

We have shown that a simple backbone constructed by the most relevant citations can well characterize the original citation network. Conversely, nontrivial information stored in the citation network can be simply extracted from its backbone. While conventional spanning trees are based on contemporary information, we demonstrated the significance of temporal dimension in citation backbones.

Specifically, we have introduced both a simple approach to identify the most relevant reference for each publication and effective measures to quantify the validity of the resulting backbone. Our results show that the essential features of hierarchy and paper clustering in the original network are well captured by our citation backbone, while this is not the case for other simple approaches. On the other hand, we showed that a resemblance to self-similarity is observed in citation backbones.

In terms of applications, the backbone can be considered as a descendant chart of research papers, which constitutes a useful basis for identifying seminal papers and paper clusters, and in general a synthetic picture of different research fields. In particular, paper classification by means of the backbone is computationally efficient when compared to the conventional clustering approaches, and provides additional information on the cluster structure besides a mere cluster label.

While we only investigated the citation network of the American Physical Society, the same approach can be readily applied to other citation networks. It would also be interesting to examine the potentials of the present approach on other directed acyclic graphs.

ACKNOWLEDGMENTS

We thank Matús Medo for fruitful discussions and comments, and Giulio Cimini and An Zeng for meaningful suggestions. This work is supported by QLectives projects (EU FET-Open Grants 231200), the Swiss National Science Foundation (Grant No. 200020-132253), the Sichuan Provincial Science and Technology Department (Grant No. 2010HH0002), and the National Natural Science Foundation of China (Grants No. 90924011 and No. 60973069). C.H.Y. is partially supported by EU FET FP7 project STAMINA (FP7-265496). We are grateful to the APS for providing us the dataset.

[1] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, 1987).
[2] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing* (Oxford University Press, Oxford, UK, 2001).
[3] P. Bak, K. Christensen, L. Danon, and T. Scalon, Phys. Rev. Lett. **88**, 178501 (2002).
[4] A. Corral, Phys. Rev. Lett. **92**, 108501 (2004).
[5] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).
[6] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, Rev. Mod. Phys. **80**, 1276 (2008).
[7] P. Chen, H. Xie, S. Maslov, and S. Redner, J. Inform. **1**, 8 (2007).
[8] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, Phys. Rev. E **80**, 056103 (2009).
[9] T. L. Griffiths and M. Steyvers, Proc. Natl. Acad. Sci. USA **101**, 5228 (2004).
[10] D. Aldous and A. G. Percus, Proc. Natl. Acad. Sci. USA **100**, 11211 (2006).
[11] T. S. Jackson and N. Read, Phys. Rev. E **81**, 021130 (2010).
[12] K.-I. Goh, G. Salvi, B. Kahng, and D. Kim, Phys. Rev. Lett. **96**, 018701 (2006).
[13] T. Nishikawa and A. E. Motter, Phys. Rev. E **73**, 065106 (2006).
[14] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning* (Springer, Berlin, 2001).
[15] S. Fortunato, Phys. Rep. **486**, 174 (2010).
[16] Z.-X. Wu and P. Holme, Phys. Rev. E **80**, 037101 (2009).
[17] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[18] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).
[19] A. K. Jain, M. N. Murty, and P. J. Flynn, ACM Comput. Surv. **31**(3), 264 (1999).
[20] C. Song, S. Havlin, and H. A. Makse, Nature (London) **433**, 392 (2005).
[21] J. S. Kim, K.-I. Goh, B. Kahng, and D. Kim, New J. Phys. **9**, 177 (2007).