SCIENTIFIC REPORTS

OPEN

Ollivier-Ricci Curvature-Based Method to Community Detection in Complex Networks

Received: 26 February 2019
Accepted: 21 June 2019
Published online: 05 July 2019

Jayson Sia, Edmond Jonckheere & Paul Bogdan

Identification of community structures in complex network is of crucial importance for understanding the system's function, organization, robustness and security. Here, we present a novel Ollivier-Ricci curvature (ORC) inspired approach to community identification in complex networks. We demonstrate that the intrinsic geometric underpinning of the ORC offers a natural approach to discover inherent community structures within a network based on interaction among entities. We develop an ORC-based community identification algorithm based on the idea of sequential removal of negatively curved edges symptomatic of high interactions (e.g., traffic, attraction). To illustrate and compare the performance with other community identification methods, we examine the ORC-based algorithm with stochastic block model artificial networks and real-world examples ranging from social to drug-drug interaction networks. The ORC-based algorithm is able to identify communities with either better or comparable performance accuracy and to discover finer hierarchical structures of the network. This opens new geometric avenues for analysis of complex networks dynamics.

Community structures are inherently found in diverse complex networks from technological, biological to social networks. As such, identifying these communities can reveal valuable information regarding the network's function, structure and organization, and vulnerability. Depending on the type of the network, these communities can represent anything from related web pages in the Internet[1], functional and chemical pathways in drug-drug interaction networks[2], to affiliations in social networks[3,4], to name a few. The community detection of complex networks is an active area of research; however, some consider this an ill-defined problem with no universally accepted definition of what constitutes a "community" nor clear guidelines in assessing its performance. As such, there have been various proposed algorithms utilizing different concepts from edge betweenness, label propagation, to graph modularity.

In this work, we propose a novel geometric approach in network community identification by using the Ollivier-Ricci curvature[5] (ORC) concept. The notion of curvature, as in Riemannian geometry, quantifies how geodesic paths converge ($ORC > 0$) or diverge ($ORC < 0$). The ORC is a coarse version of this concept, and its application to graphs reveals local topological structure and geometry via optimal transport. The ORC captures the notion of network flows of shortest paths via the Wasserstein's distance formulation wherein a negatively curved edge is a "bottleneck", along which traffic is intense in a scheme that minimizes the "cost" of transferring "commodities", while, positively curved edges contribute to transport of "commodities" along with many other edges. Thus, positively curved edges are "well connected", since none of them are essential for the proper transport operation; therefore, positively curved edges naturally form a "community". On the other hand, negatively curved edges could be interpreted as "bridges" between communities and cutting them would isolate the network flow between communities. In this context, a "community" is defined as a robust transport of information within the community. Robust means that if some edges are cut information is still going to flow. As an example, for a social network, one hears news not just from one single source but from many different sources. On the other hand, information transfer across communities is more problematic since it relies on these "highway" links that could fail if these connections were removed. The ORC is also recently being applied as a tool in various research areas such as in wireless networking[6,7], quantum computation[8,9] as well as robustness analysis of complex networks[10,11].

Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089, USA. Correspondence and requests for materials should be addressed to J.S. (email: jsia@usc.edu)

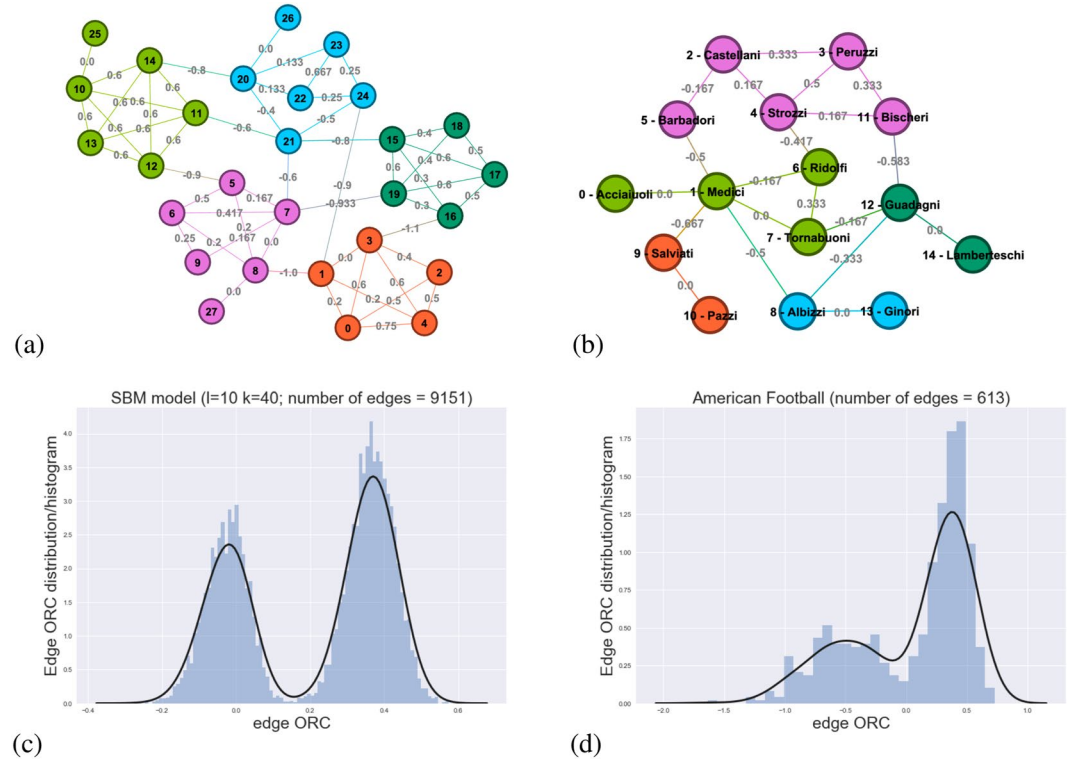

Figure 1. (**a**) An artificial network generated from the stochastic block model (SBM) with the following parameters: size per community $k = 5$, number of communities $l = 5$, intra-cluster probability $p_{in} = 0.8$ and inter-cluster probability $p_{out} = 0.05$, with added extra leaf nodes (nodes: 25, 26, 27) to illustrate edges with zero curvature, and (**b**) the Florentine family network with edge ORC values shown for both networks. Edge ORC histograms and distribution fit shown for larger networks: (**c**) 800-node SBM artificial network ((k, l, p_{in}, p_{out}) = (40, 10, 0.7, 0.05)) and (**d**) the American Division IA college football games during regular season of Fall 2000.

Results

Ollivier-ricci curvature as a natural metric to discover hierarchies in graphs. The Ollivier-Ricci curvature (ORC) captures two fundamental properties of the structure of complex networks: First, the ORC associated with each edge of the network encodes its shortest path characteristics[6]. Second, the ORC provides information about the frequency of triangles, characterized by the clustering coefficient, within a neighborhood of two adjacent vertices[12,13] (for mathematical details on how we estimate the ORC for a weighted graph see Methods and Section 2 in Supplementary Information). Starting from these premises, in this work, we aim to address the following questions: Can the ORC help us discover the underlying hierarchical functional characteristics of a complex network? Can the ORC curvature provide algorithmic hints towards solving the hard problem of community identification?

To address these questions, we consider an artificial complex network obtained through the stochastic block model (SBM)[14] as seen in Fig. 1a and a real-world social network (the Florentine family[15] network) as seen in Fig. 1b. For completeness, Fig. 1c,d illustrate the distribution of edge ORCs for larger networks, one artificial and one real-world network. From Fig. 1a, we make the following observations: (i) edges within a cluster have positive ORC values; (ii) peripheric nodes have zero ORC values; and (iii) edges between clusters have negative ORC values. Range of values for the edge ORC is $[-\infty, 1]$. For networks with clear community structures, especially seen in Fig. 1c,d, the distribution of edge ORC values is clustered into two regions – one region of positive and another region of negative edge ORCs. The high concentration of positive edge ORC values corresponds to the intra-community edges while the concentration of negative edge ORC values corresponds to the inter-community edges. Along the same lines and as can be seen from Fig. 1(a), the positively curved edges form a tight-knit neighborhood of nodes. In contrast, negatively curved edges represent links between tightly connected neighborhoods. These observations suggest that a community identification would proceed by incrementally removing negatively curved edges, and the absence of such negatively curved edges is a natural stopping criterion. We describe the pseudocode and formal analysis of this ORC-based CI algorithm in Methods section and Section 2 of Supplementary Information.

ORC-based CI detects communities in artificially generated networks. To investigate the accuracy of the proposed ORC-based community identification (CI) algorithm (see Methods Section), we consider a set of artificially constructed complex unweighted networks using the stochastic block model (SBM)[14]. The SBM consists of four parameters, i.e., k represents the size of a community, l denotes the number of communities, and

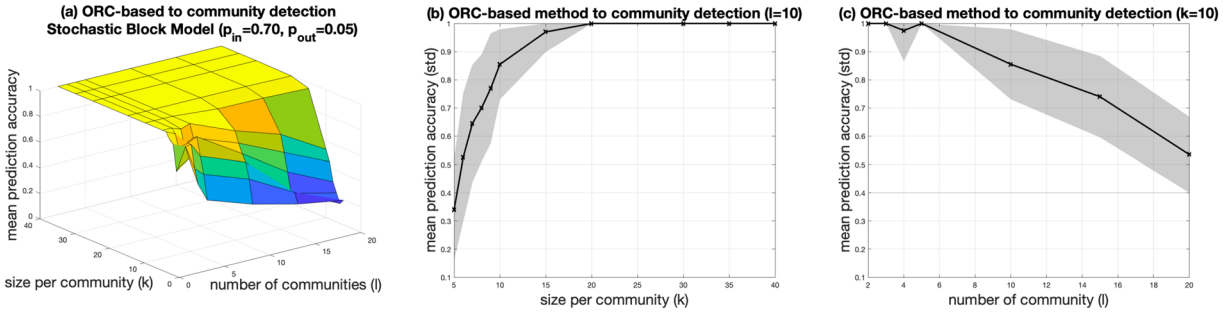

Figure 2. (**a**) Surface plot of the mean prediction accuracy of the ORC-based community detection method for several artificially generated networks using the stochastic block model (SBM) with intra- and inter-community edge wiring probability settings: $p_{in} = 0.70$, $p_{out} = 0.05$. Prediction mean accuracy with standard deviation bands for (**b**) varying size per community (k) parameter for $l = 10$ number of communities, and (**c**) varying number of communities (l) parameter for $k = 10$ size per community.

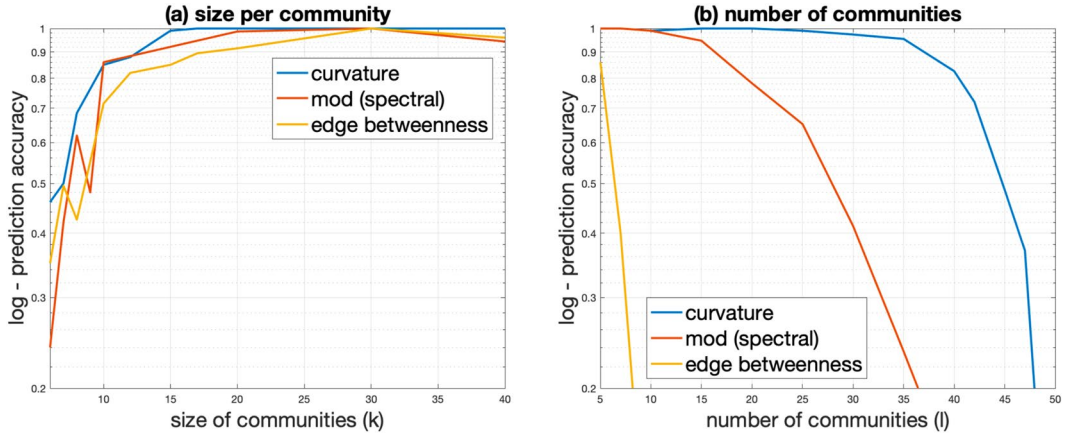

Figure 3. Prediction accuracy of the ORC-based, LEM and EB CIs on SBMs based on varying size per community and number of communities parameters. Intra- and inter-community edge wiring probabilities are set to $p_{in} = 0.7$ and $p_{out} = 0.05$, respectively. (**a**) Prediction accuracy as a function of size of communities for $l = 10$ number of communities. (**b**) Prediction accuracy as a function of number of communities for $k = 20$ size per community.

p_{in} and p_{out} are the probability of creating intra- and inter-community edges between any two nodes. Thus, SBM generates a complex network with user-defined community sizes and labels each node accordingly.

Figure 2a shows the mean prediction accuracy surface plot of the ORC-based CI algorithm applied to artificially generated graphs obtained from the SBM with $p_{in} = 0.7$ and $p_{out} = 0.05$, respectively. In this study, we vary the number of communities from 2 to 20 and the community size from 5 to 40. From Fig. 2b, we can observe that for a fixed number of communities (l), the accuracy improves from 33% for small community sizes to 100% for large community sizes. On the other hand, for fixed size per community, there is a degradation in prediction accuracy as the number of communities is increased. For completeness, Fig. 2b,c show the accuracy and its confidence interval as a function of the community size and the number of communities, respectively.

We compare the ORC-based CI algorithm with the modularity-based Leading Eigenvalue method[16] (LEM) and edge betweenness[4] (EB) based CI algorithms. Figure 3 shows the prediction accuracy for the ORC-, LEM- and EB-based CI algorithms. The accuracy of each algorithm measures the percentage of correctly identified communities from the set of ground truth communities. More precisely, when a set of nodes identified as a community matches all the members of a set from the list of ground truth communities, this is considered a correctly identified community. The prediction accuracy is obtained as a percentage of the correctly identified communities over all ground truth communities. In addition, the results are presented in the context of the SBM detectability regime[17] which describes the phase transition in the detectability of communities subject to the chosen SBM parameters. For cases when the chosen SBM parameters lie in the undetectable regime, the SBM generated graph is indistinguishable from a random generated graph and community detection is impossible.

As we vary the size per community while keeping the number of communities and the probability of intra- and inter-community edges between any two nodes (p_{in} and p_{out}) constant (see Fig. 3a), we observe the accuracy of the ORC-based CI improves significantly for networks with more than 7 communities. The low accuracy for sizes k less than 7 is due to the high probability to merge two small-sized communities caused by the addition of inter-community edges between the two communities. This observation falls close to the theoretical detectable

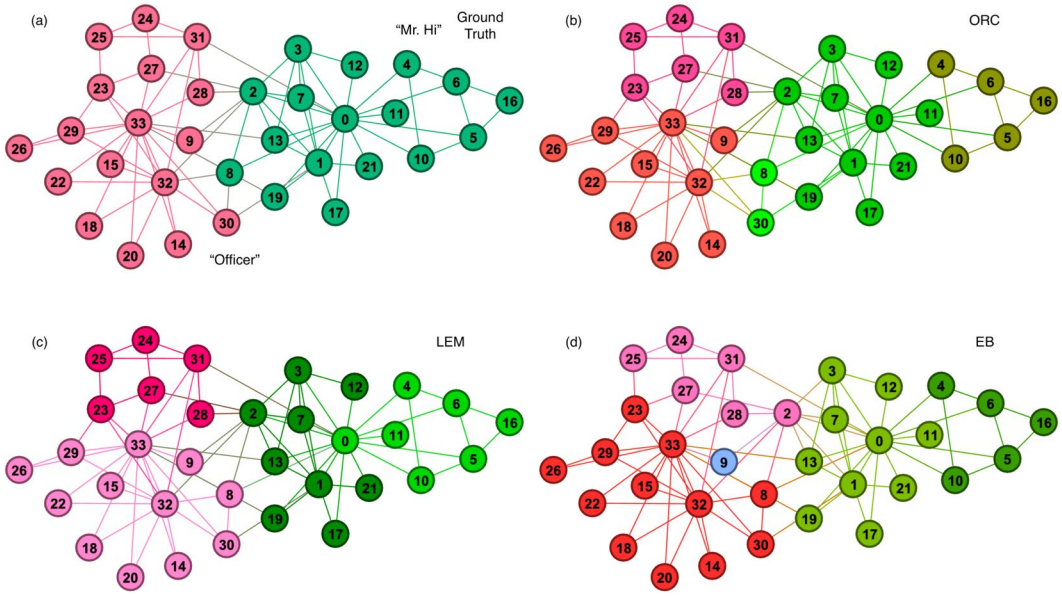

Figure 4. Zachary's Karate Club. (**a**) Shows the ground truth divided between two communities: 'Officer' (in red) and 'Mr. Hi' (in green). (**b**–**d**) Show the communities detected (color-coded) using the ORC-, LEM- and EB-based CI methods.

regime of $k^* \geq 3$ for varying sizes per community k and fixed $l = 10$, $p_{in} = 0.7$ and $p_{out} = 0.05$. For sizes of community less that 15, the ORC-based CI performs better compared to both LEM and EB methods. There is also a slight degradation in accuracy for the LEM- and EB-based CIs after community size greater than 30.

As we vary the number of communities but keeping the size per community and the probability of intra- and inter-community edges between any two nodes (p_{in} and p_{out}) constant (see Fig. 3b), we observe the accuracy of the ORC-based CI degrades significantly for networks with more than 35 communities. In contrast, the LEM demonstrates worse performance as its accuracy degrades significantly for more than 15 communities. Both, the ORC- and LEM-based CI reach 50% accuracy for more than 30 communities. The EB-based CI performs the worst out of the three CI algorithms. For varying number of community l and fixed $p_{in} = 0.7$ and $p_{out} = 0.05$, the SBM detectable regime lies in $l^* \leq 156$.

In short, the ORC-based CI can detect the communities for SBM generated networks even when the communities are not densely connected and even for small community sizes. However, the accuracy degrades when the probability of intra-community edge wiring is less than 0.6 ($p_{in} < 0.6$) and the probability of inter-community edge wiring is greater than 0.04 ($p_{out} > 0.04$), for fixed size per community and number of communities ($k = 20$ and $l = 30$). In comparison, the accuracy of the LEM-based CI degrades when the probability of intra-community edge wiring is less than 0.7 ($p_{in} = 0.7$) and the probability of inter-community edge wiring is greater than 0.03 ($p_{out} > 0.03$). For the chosen $(k, l) = (20, 30)$, the detectability regime lies at $p_{in}^* > 0.35$ for constant $p_{out} = 0.05$, and $p_{out}^* < 0.17$ for constant $p_{in} = 0.7$. Simulation results show that the ORC-based CI provides better or comparable accuracy with the LEM-based CI (see Section 3 of Supplementary Information for a complete discussion and illustration of results which include the prediction accuracy with respect to the intra- and inter-community wiring probabilities p_{in} and p_{out} in the context of the theoretical SBM detectability limit[17]).

ORC-based method applied to real-world networks. Next, we test the proposed ORC-based CI algorithm to a few real-world network data: Zachary's Karate Club, American Football games, Political Blogosphere, and DrugBank drug-drug interaction network. We compare the obtained results to other popular established community detection algorithms.

Zachary's karate club. Figure 4 shows the network visualization for the Zachary's karate club which is a traditional dataset and a standard benchmark in community detection[3]. The network contains 34 nodes (members) and 78 edges (friendships) with an average degree of 4.6. Figure 4a shows the ground truth divided between two communities: 'Officer' (in red) and 'Mr. Hi' (in green). Figure 4b shows the 5 communities identified by the ORC-based method. Although there are more communities identified, two sub-communities (marked in shades of red) match correctly with the ground-truth subgraph labeled as 'Officer' and two sub-communities (marked in shades of green) match correctly with the ground-truth subgraph labeled as 'Mr. Hi'. The fifth community identified (nodes 8 and 30) neither falls correctly into either communities and thus can be considered as a classification error. From the results, the ORC-based method identifies finer hierarchical structures in the network. More precisely, it identifies further subdivisions within the known truth labeled communities. Figure 4c,d, on the other hand, show that the LEM- and EB-based CI methods identify 4 and 5 communities, respectively. Both assign node 8 as part of the larger community associated with 'Officer' which is a classification error. In addition, the EB-based CI has misclassified node 2, apart from labeling node 9 as a separate single-node community.

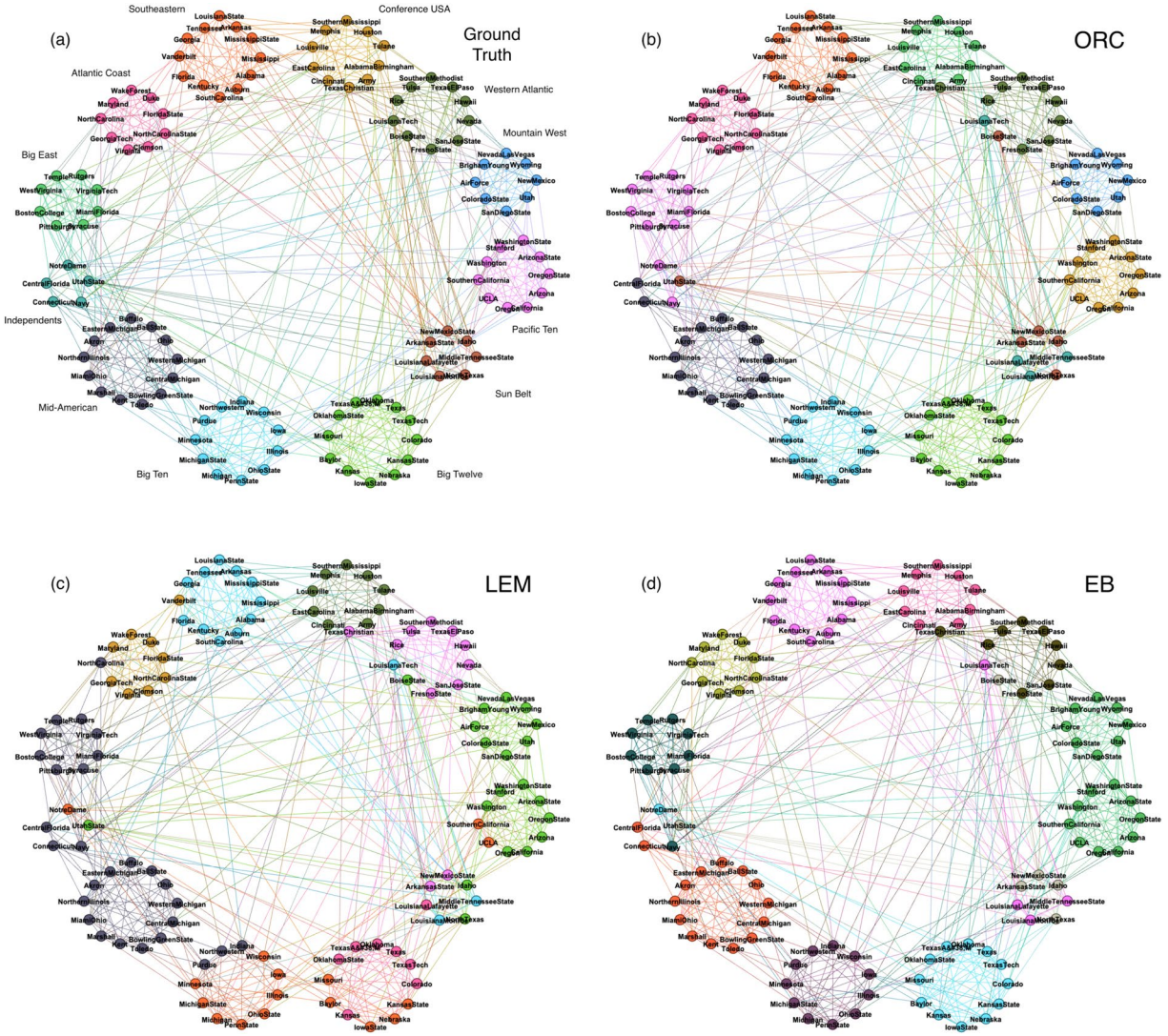

Figure 5. Network of American football games (Division IA) during regular season Fall 2000. (**a**) Each node represents a college color-coded according to its football conference membership (ground truth). (**b**–**d**) Nodes are color-coded according to the communities identified by the (**b**) ORC-, (**c**) LEM-, and (**d**) EB-based CI methods, respectively.

American college football games 2000. Figure 5a shows the network visualization for the network of American football games between Division IA colleges during regular season of Fall 2000[4]. The colleges are grouped together and color-coded according to their football college conference memberships (e.g. Pac-10, Big-12, etc.). The twelve college football conferences are considered as the ground truth communities. The network contains 115 nodes (colleges) and 613 edges (games) with an average degree of 10. The ORC-based method is able to detect twelve college communities which is also the same number as compared to the list of ground truth football conferences. As seen in Fig. 5b, the ORC-based CI is able to assign all members of 8 out of 12 communities together. If we consider per-node classification accuracy, the ORC-based method incurred 11 misclassified nodes (9.5% misclassification). On the other hand, as seen in Fig. 5c,d, the LEM- and the EB-based CIs are able to identify 8 and 10 college communities, respectively, which are both less than the number of ground truth communities. The per-node classification accuracies are 41 and 19 misclassified nodes translating to 35.7% and 16.5% misclassifications, respectively. In addition, the EB-based CI has identified both ground truth communities "Mountain West" and "Pacific Ten" as one community while the LEM-based CI, in addition to this, has also merged both ground truth communities "Big East" and "Mid-American" as one. Thus, among the three CI methods used, the ORC-based CI performed the best in terms of both correct identification of ground truth communities and per-node community assignments.

Interesting to note that ground truth communities labelled as 'Independents' (Notre Dame, Utah State, etc.) and 'Sun Belt' (Arkansas State, Idaho, etc.) have no clear community assignments according to three CI methods used. Closer inspection shows that there are very low intra-community links among the members, thus it can be argued that the "Independents" and "Sun Belt" are not really a community in the strict-sense. As another example,

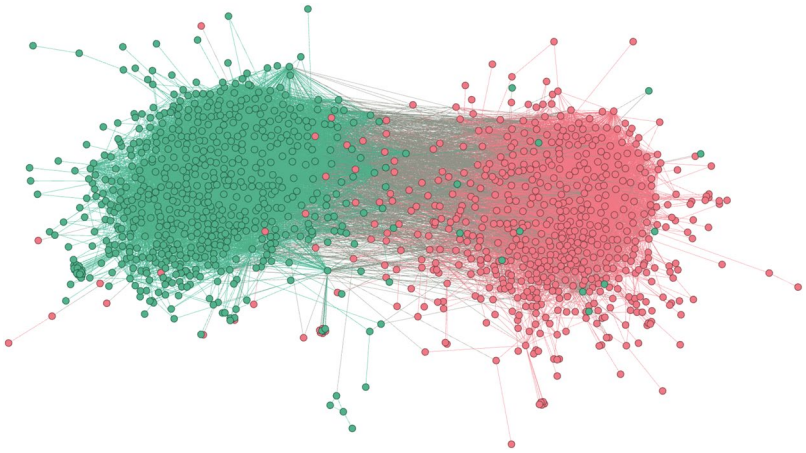

Figure 6. Political Blogosphere 2005 (Nodes = 1222, edges = 16714, average degree = 27.35). Network visualization shows the ground truth communities (green - right/conservative, red - left/liberal).

Texas Christian is misclassified to "Western Athletic" by all three CI methods but the ground truth assignment is "Conference USA". Visual inspection of the network shows that Texas Christian has more edge connections (games) with the colleges belonging to the "Western Athletic" compared to those belonging to "Conference USA" which explains the misclassification.

Political blogosphere 2005. Figure 6 shows the network visualization for the 2005 Political Blogosphere[1] which is a network of blog directories labelled according to a blog's political leaning (left/liberal or right/conservative). Considering only the largest connected component, the network contains 1222 nodes (web URLs) and 16714 edges with an average degree of 27.35. Node truth labels indicate political leanings (0 - left/liberal (52.05%); 1 right/conservative (47.95%)). Links between blogs were automatically extracted from a crawl of the front page of the blog. Data on political leaning comes from blog directories with some blogs labeled manually based on the incoming and outgoing links and posts around the time of the 2004 presidential election. The ORC-based method is able to find 146 communities. The top two largest identified communities have sizes 448 and 400 while the rest of the 144 communities have sizes less than 1% of the network size (size 10). Since we know that there are only two ground truth communities, we did preferential attachment of the smaller-sized communities to either of the two largest components based on the inter-community ORC. We identified communities with size 662 (54.17%) for left/liberal and 560 (45.83%) for the right/conservative after preferential attachment of the smaller-sized communities. On the other hand, the LEM-based CI is able to identify two communities with community sizes 677 (55.4%) for left/liberal and 545 (44.6%) for right/conservative. For this dataset, the ORC-based CI performed better than the LEM-based CI in identifying the communities based on the binary ground truth labels.

Drug-drug interaction. We also analyze the community structure of drug-drug interaction networks obtained from DrugBank 4.1 database[18] which was demonstrated as an efficient method for drug repositioning[2]. The network visualization is shown in Fig. 7. Considering only the largest connected component from the dataset, the network contains 1162 nodes (drugs) and 11685 edges (drug interactions) with an average degree of 20. Figure 7a,b show the communities identified by the ORC- and LEM-based CIs, respectively. The ORC-based CI initially identifies 120 communities of which 101 of them have sizes smaller than 1% of the network size. Setting this 1% threshold size limit, we can apply preferential attachment based on the inter-community ORC to merge back these small-scale communities to the larger identified communities. The 19 communities identified by the ORC-based CI are color-coded as shown in Fig. 7a. On the other hand, the LEM-based CI is able to identify 4 communities as shown in Fig. 7b. Comparing the results from Udrescu's paper[2], the ORC-based CI matches closely the topological clusters generated based on the energy-model layout algorithm Force Atlas 2, albeit the ORC-based CI identifying 19 communities compared to the 9 labeled topological clusters. Visual inspection of Fig. 7a shows that the ORC-based CI identifies sub-clusters within the topological clusters.

Discussion

The study of community identification in complex networks is an important and challenging open area of research. Many diverse systems can be represented into networks many of which have inherent community structures. The network abstraction offers a simpler way of looking at a system's individual elements as nodes and their interactions as edges. In addition, the inherent community structures within networks convey information regarding the system's function, hierarchy and organization. In the past decade, there have been numerous algorithms[4,16] proposed to solve the problem of community identification each having its own advantages and limitations. Our proposed ORC-based approach offers a new way to tackle the community identification problem by utilizing the geometric concept of curvature applied to discrete graphs.

From simulations, the ORC-based CI is able to identify communities from the SBM-generated artificial networks with either better or comparable performance accuracies as compared to the LEM- and EB-based CIs. In

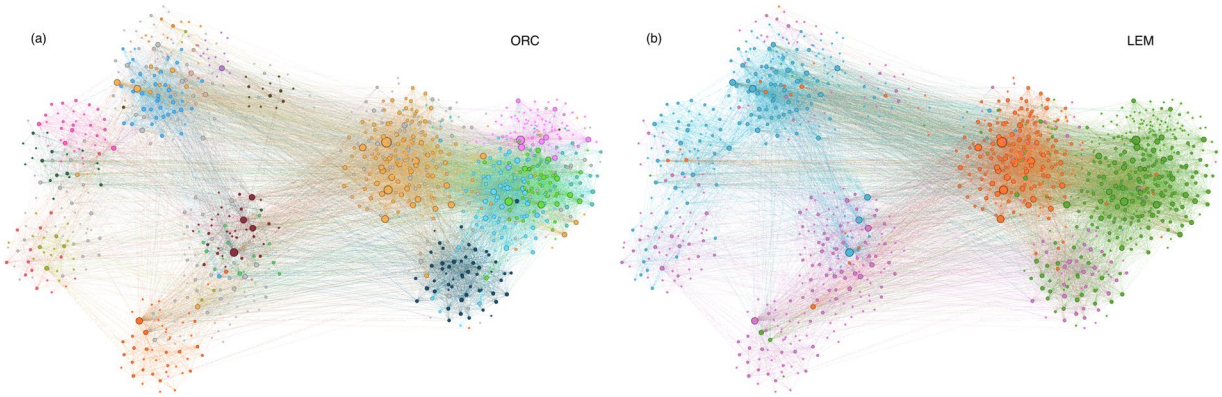

Figure 7. Drug-drug interaction dataset obtained from the DrugBank v4.1 database. Node layout is based on generated topological clusters using Gephi[31] software with energy-model layout algorithm Force Atlas 2. Nodes are color-coded according to the communities identified by the (**a**) ORC- and (**b**) LEM-based CI methods, respectively.

addition, we also observe that the ORC-based CI performs well in identifying community structures for diverse real-world networks ranging from social to drug-drug interaction networks. For example, the ORC-based CI results for the drug-drug interaction network matches closely the identified topological clustering as compared to the LEM-based modularity method. As seen from the American football and the drug-drug interaction network examples, the LEM-based CI tends to underestimate the total number of communities. One limitation of modularity-based methods is that it has a resolution limit that may prevent it from detecting clusters that are comparatively small compared to the graph as a whole[19]. Contrary to this, the ORC-based CI tends to see the finer subdivisions in the network structures (i.e. identifying "communities within communities") based on the local topology as quantified by the edge Ollivier-Ricci curvature. Thus, the ORC-based CI tends to identify more communities compared to the number of ground truth communities.

Since the proposed ORC-based CI stems from the idea of successive removal of negatively curved edges, the ORC-based CI will not perform well for networks that have an almost tree-like topology (i.e. graphs that have very few cycles/triangles). This is because tree-like networks have negatively curved edges forming a majority. As a consequence, the ORC-based CI will divide the network into several small communities, thus highly overestimating the number of communities. There are a couple of ways to see this limitation: First, one can argue that tree-like community structures are not considered as a community in the strict-sense since the community members have very few connections among one other. Thus, labelling such communities is considered an ill-defined problem. Second, if the number of communities is known in the first place, preferential attachment heuristics can be applied to re-attach these small-scale communities back to form larger communities.

In conclusion, the ORC-based CI offers a novel alternative solution to the problem of network community identification. Since the algorithm utilizes the geometric concept of network curvature, the ORC-based CI performs particularly well for networks with internally densely-connected community structures. For community structures that are sparsely connected, the ORC-based CI will tend to overestimate the number of communities as it identifies the finer community structure. Preferential attachment heuristics can be applied to merge back these small-scale communities to the larger communities. This preferential attachment portion of the algorithm can be further explored as a future work especially for CI problems with both known and unknown number of communities.

Methods

ORC-based community detection algorithm. We propose the following community detection algorithm which utilizes the concept of Ollivier-Ricci curvature on graphs. The algorithm can be divided into the following steps: (1) calculate the ORC for all edges in the network, (2) remove the most negative ORC edge, (3) re-calculate the edge ORC only for those affected nodes/edges due to prior edge removals, (4) check if all edge

Algorithm 1. Ollivier-Ricci Curvature based method for Community Detection

Input : A graph object $G(V, E, \rho)$ with a list of nodes, V, and a list of edges, E.
Output : A graph object $G'(V, E, \rho)$ similar to G but with additional node properties indicating community label.

$G' \leftarrow G$ with the Ollivier-Ricci curvature calculated for all edges;
while *there exists a negative edge curvature in* G' **do**
    Remove the most negatively curved edge;
    Re-calculate the Ollivier-Ricci curvature for the affected existing edges in G';
end
`PreferentialAttachment` (G',*number_of_communities, minimum_community_size*) ;
Label each node with a unique community label according to its membership to a particular graph component;
return G'

curvatures are non-negative, otherwise repeats steps 2 and 3 until condition is satisfied, and (5) perform preferential attachment of isolated graph components if either the number of communities or the minimum accepted community size is known. The pseudo-code of the algorithm is seen in Algorithm 1.

Complexity. The time complexity of the proposed ORC-based community detection algorithm boils down to the calculation of the edge ORC of the network. The time complexity to compute the ORC for each edge is essentially the Wasserstein distance computation complexity based on linear programming. Practical run time complexity using network (transportation) simplex algorithm[20] was shown to be super-cubic. Interior-point or Orlin's algorithms have complexity of $O(V^3 \log V)$, with V as the total number of vertices in the Wasserstein distance sub-problem[21,22] (Note that V depends on twice the average degree of the network typically with $V \ll N$ and $V \ll E$). In the worst case, cycling through each network edge and re-calculating all existing affected edges lead to $O((EV) \cdot V^3 \log V))$. Strategies can be utilized to improved the computation complexity of the proposed algorithm either via a wavelet EMD approximation[21] of the Wasserstein distance or an ORC bounds analysis[12,23]. The Wasserstein distance computation can be improved from $O(V^3 \log V)$ to $O(V)$ via the wavelet EMD approximation leading to an overall time complexity of $O(EV^2)$ for the proposed algorithm. More information regarding the time complexity of the proposed ORC-based CI and its code implementation are provided in Section 2 of the Supplementary Information.

Foundation of differential geometry. Fundamental in the process of extending geometry in the Euclidean plane to geometry on a surface $\mathcal{S} \subset \mathbb{R}^3$ is the intuitive idea of projecting the ordinary derivative $\frac{d}{dt}X(c(t))$ of a tangent vector field X, defined along a curve c, on the tangent space to the surface, leading to the concept of Levi-Civita connection

$$\nabla_{\dot{c}} X := P_{T_{c(t)}\mathcal{S}}\left(\frac{d}{dt}X(c(t))\right) \in T_{c(t)}S.$$

The covariant derivative $\nabla_C X$ of the vector field X along the vector field C (not necessarily the tangent to a curve) in a Riemannian manifold $\mathcal{M}$ is a formalization of the intuitive geometric concept of restricting the differential to the tangent space, subject to the additional conditions of symmetry, $\nabla_C X = \nabla_X C$, linearity relative to C, the product rule relative to scalar multiplication of X and compatibility with the Riemannian metric, viz., $\frac{d}{dt}\langle X(c(t)),\, Y(c(t))\rangle = \langle \nabla_{\dot{c}} X,\, Y\rangle + \langle X,\, \nabla_{\dot{c}} Y\rangle$.

A vector field X is said to be *parallel to itself* along the curve c: $[0,\, 1] \to \mathcal{M}$, if it satisfies the partial differential equation $\nabla_{\dot{c}} X = 0$. Under such conditions, $X(c(1))$ is said to be a *parallel displacement* of $X(c(0))$. This formal definition calls into question by how much this parallel displacement differs from the ordinary Euclidean one. A nonvanishing curvature is precisely symptomatic of such discrepancy. But the immediate problem is that $X(c(0))$ and $X(c(1))$ lives in different tangent spaces and are difficult to compare. One way to go around this difficulty—challenged by the Ollivier[5] concept of curvature—is to bring $X(c(1))$ back to $T_{c(0)}\mathcal{M}$ by another parallel displacement along an extension of c to a closed curve. To somewhat simplify the problem without sacrificing generality in our Ollivier-Ricci curvature objective, assume the curve c and the vector field X live in a 2-dimensional tangent bundle span$\{X, Y\}$. Then

$$\angle(X(c(1)),\, X(c(0)) = \text{Area}\,(c) K(X,\, Y), \tag{1}$$

where $K(X,\, Y)$ is the sectional curvature, a curvature where the parallel displacement is restricted to a 2-dimensional facet. Precisely,

$$K(X,\, Y) = \frac{\langle R(X,\, Y)X,\, Y\rangle}{\|X\|^2 \|Y\|^2 - \langle X,\, Y\rangle^2},$$

where

$$R(X,\, Y) = \nabla_Y \nabla_X - \nabla_X \nabla_Y + \nabla_{[X,Y]}$$

is the fundamental curvature operator.

Connection with wireline networks and diffusion processes. Wireline networks in general send packets along optimal paths, along *geodesics* in Riemannian language. Note that a geodesic is only locally length $\ell(\gamma) = \int_\gamma ds$ optimal, as formally the geodesic is defined such that its tangent is parallel to itself, $\nabla_{\dot{\gamma}} \dot{\gamma} = 0$, where the geodesics is parameterized by arc length and $\dot{\gamma} := \frac{d\gamma(s)}{ds}$. Motivated by network outages where optimal paths have to be quickly recomputed, the nominal geodesic γ is embedded in a family of geodesics, γ_p, $p \in (-\varepsilon, +\varepsilon)$ with $\gamma_0 = \gamma$. The *Jacobi field* $J(s) := \frac{d}{dp}\gamma_p(s)\Big|_{p=0}$, quantifying the variation of geodesics, satisfies the equation

$$\nabla_{\dot{\gamma}} \nabla_{\dot{\gamma}} J + K(J,\, \dot{\gamma}) J = 0. \tag{2}$$

Under uniform curvature K, it is convenient to search a solution of the form $J(s) = j(s) W(s)$, where $W(s)$ is orthogonal to $\gamma(s)$, in which case

$$\frac{d^2}{ds^2} j(s) + K j(s) = 0. \tag{3}$$

Clearly, if $K<0$, geodesics are diverging, an observation that lies at the foundation of congestion in wireline Gromov hyperbolic networks[24].

Other processes of the diffusion type, that is, such processes as heat diffusion and Heat Diffusion wireless networking[6,7,25–28] involving the Laplace operator, do not "diffuse" along geodesics, but rather follow some thermodynamical-like processes, where the heat kernel exposes the curvature in its Ricci format. The *Ricci curvature* $\mathrm{Ric}(X)$ is the average of $K(X, Y)$ over all facets $\mathrm{span}\{X, Y\}$ containing X.

Note the fundamental difference between wireline-like networking and diffusion. Wireline networking involves large-scale optimal paths, whereas wireless networking in both its backpressure and Heat Diffusion implementations is driven by strictly local queue backlogs, in the same way as heat diffusion is driven by a strictly local temperature gradient.

Towards ollivier-ricci curvature. Contrary to what is usually done, here, we attempt to define curvature by reference to different tangent spaces, one centered at $\gamma(0)$, the other at $\gamma(\varepsilon)$. Consider two δ-radius balls $B_{\gamma(0)}, B_{\gamma(\varepsilon)}$. We establish a correspondence between the two balls as follows: Consider $x\in B_{\gamma(0)}$ along with $X=\exp^{-1}_{\gamma(0)}(x)$. Displace X parallel to itself along γ from $\gamma(0)$ to $\gamma(\varepsilon)$ to obtain Y. Define $y=\exp_{\gamma(\varepsilon)}(Y)$. This establishes the correspondence T: $x\mapsto y$. To introduce a *transport* idea, the ball $B_{\gamma(0)}$ is endowed with a probability measure μ_0 and $d\mu_0(x)$ is transported to $y=T(x)$ along a geodesic arc $[x, y]$ of length equal to the distance $d(x, y)$.

Invoking the Jacobi field (2 and 3), the distance $d(x, T(x))$ along the "perturbed" geodesic $[x, y]$ and how it relates to the distance $d(\gamma(0), \gamma(\varepsilon)=\varepsilon)$ along the "nominal" geodesic depends on the sectional curvature $K(X, \dot\gamma)$. Therefore, the cost of the transport

$$C(T)=\int_{B_{\gamma(0)}} d(x, T(x))d\mu_0(x), \tag{4}$$

since it involves an integral over all $x\in B_{\gamma(0)}$, tacitly involves an integral over all tangent vectors $X\in T_{\gamma(0)}B_{\gamma(0)}$ and as such averages $K(X, \dot\gamma)$ over all X to yield the Ricci curvature $\mathrm{Ric}_{\dot\gamma(0)}(\mathcal{M})$.

In 0-curvature, the distance $d(x, T(x))$ is independent of x and therefore the transport cost is $d(\gamma(0), \gamma(\varepsilon)=\varepsilon)$. It remains to see how this distance is affected by the curvature. Define $d\theta(s)$ to be the elementary angle swept by the normal $W(s)$ to the geodesic under an elementary move ds along such geodesic. Then

$$d(x, T(x))=\varepsilon+\int_0^{\varepsilon} j(s)d\theta(s).$$

$j(s)$ is the distance between the nominal and perturbed geodesics measured along the normal to the nominal geodesic; using (3), it is evaluated as

$$\begin{aligned} j(s) &= \delta\cosh(\sqrt{-K}s)-\frac{\varepsilon\delta}{2}\sqrt{-K}\sinh(\sqrt{-K}s)\\ &\approx \delta\cosh(\sqrt{-K}s). \end{aligned}$$

Next, we apply (1) to the closed path made up with $\dot\gamma_0(s)ds$, $j(s+ds)W(s+ds)$, $-\dot\gamma_\delta(s+ds)ds$ and $-j(s)W(s)$. Noting that the left-hand side of (1) is the full discrepancy angle around the closed path while we only need the discrepancy along the nominal geodesic, we get

$$d\theta=\frac{1}{2}d\,\mathrm{Area}(j, ds)\sqrt{-K} \tag{5}$$

$$=\frac{1}{2}j(s)ds\sqrt{-K}. \tag{6}$$

Putting everything together and after an elementary integration, it is found that

$$d(x, T(x))\approx\varepsilon\left(1-\frac{1}{2}K\delta^2\right),$$

an estimate consistent with that of [5, Prop. 6, Sec. 8].

The above estimate was derived nominally in a negatively curved manifold, but redeveloping the same argument with ordinary trigonometry rather than hyperbolic trigonometry would validate it in positively curved spaces.

The above clearly indicates that in negative curvature, the transportation cost from x to $T(x)$ is larger than along the nominal geodesic. In positive curvature, the x to $T(x)$ cost is smaller than along γ.

To summarize:

$$\begin{aligned} \mathrm{Ric}_{\dot\gamma(0)}(\mathcal{M})<0 &\Leftrightarrow \int_{B_{\gamma(0)}} d(x, T(x))d\mu_0(x)>d(\gamma(0), \gamma(\varepsilon)),\\ \mathrm{Ric}_{\dot\gamma(0)}(\mathcal{M})=0 &\Leftrightarrow \int_{B_{\gamma(0)}} d(x, T(x))d\mu_0(x)=d(\gamma(0), \gamma(\varepsilon)),\\ \mathrm{Ric}_{\dot\gamma(0)}(\mathcal{M})>0 &\Leftrightarrow \int_{B_{\gamma(0)}} d(x, T(x))d\mu_0(x)<d(\gamma(0), \gamma(\varepsilon)). \end{aligned}$$

From riemannian manifolds to graphs. On a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ endowed with a distance $d(\cdot,\cdot)$, we need to emulate the Riemannian manifold environment. We identify an edge ij of the graph with the geodesic $\gamma([0, \varepsilon])$ and the graph theoretic neighborhoods $\mathcal{N}_i$, $\mathcal{N}_j$ of i and j with the balls $B_{\gamma(0)}, B_{\gamma(\varepsilon)}$ centered at $\gamma(0)$, $\gamma(\varepsilon)$. Discrete probabilities μ_i, μ_j on $\mathcal{N}_i$, $\mathcal{N}_j$ are obvious substitutes for the measures μ_0, μ_ε on the balls $B_{\gamma(0)}, B_{\gamma(\varepsilon)}$.

The difficulty is to emulate the Riemannian connection resorting only to the graph theoretic distance, or at the very least redefine the cost $C(T)$ in (4) in a way that does not involve parallel displacement. Proceeding from

$$C = \inf_{T: B_{\gamma(0)} \to B_{\gamma(\varepsilon)}} \int_{B_{\gamma(0)}} d(x, T(x)) d\mu_0(x),$$

where T is restricted to be one-to-one, the graph theoretic emulation of the above is

$$\overrightarrow{C}_{\mathcal{G}} = \min_{\mathcal{N}_i \ni k \mapsto \ell \in \mathcal{N}_j} \sum_{jk \in \mathcal{N}_i} d(k, \ell) \mu_i(k)$$

In this case, because the cardinalities of $\mathcal{N}_i$ and $\mathcal{N}_j$ might not be the same, the mapping $k \mapsto \ell$, while one-to-many, could be many-to-one. As such, the formula lacks symmetry and cannot be used as a Wasserstein-like distance. To remedy this situation, we introduce a *transference plan* $\xi^{ij}(k, \ell)$ as a substitute for the many-to-many mapping $k \mapsto \ell$, with the added generality that only a piece $\xi^{ij}(k, \ell)$ of $\mu_i(k)$ is transferred to ℓ. The above formula hence becomes

$$C_{\mathcal{G}} = \min_{\xi^{ij}(k,\ell)} \sum_{k \in \mathcal{N}_i, \ell \in \mathcal{N}_j} d(k, \ell) \xi^{ij}(k, \ell)$$

with of course the consistency conditions

$$\sum_{\ell \in \mathcal{N}_j} \xi^{ij}(k, \ell) = \mu_i(k), \qquad \sum_{k \in \in \mathcal{N}_i} \xi^{ij}(k, \ell) = \mu_j(\ell).$$

The curvature concept that emanates from this cost ($C_{\mathcal{G}} > (<) \varepsilon \Leftrightarrow \mathrm{Ric} < (>) 0$) is very local, around an edge, in contradiction with the global Gromov concept. This explains why such concept appears the correct one to anticipate performance of backpressure and Heat Diffusion protocols on wireless networks[6,7].

Ollivier-ricci curvature on complex networks. The proposed community detection algorithm utilizes the coarse Ricci curvature, referred to as Ollivier-Ricci curvature, in its version designed for complex networks. Since the Ricci curvature involves a privileged direction ($\dot{\gamma}(0)$ on $\mathcal{M}$, edge ij on $\mathcal{G}$), it incorporates a generic concept of *flow*. In the Riemannian model, $\mathrm{Ric}_{\dot{\gamma}(0)} < 0$ means "heavy" flow, in the sense that the least cost transport of probability mass takes the geodesic γ path rather than being distributed along the perturbed geodesics. In the graph/network context, the ball of mass around i is the set of neighbors of i (same for j). Similarly, the idea is to find the best way to transfer the ball of mass around the vertex i to that around the vertex j.

Consider a weighted graph $((\mathcal{V}, \mathcal{E}), \rho)$. On this graph, over each vertex i, we define a probability measure on $\mathcal{N}_i := \{k \in \mathcal{V}: ik \in \mathcal{E}\}$ as follows:

$$\begin{aligned} \mu_i(k) &= \frac{\rho_{ik}}{\sum_{k \in \mathcal{N}(i)} \rho_{ik}}, \qquad \text{if } ik \in \mathcal{E} \\ &= 0 \quad \text{otherwise} \end{aligned}$$

The Ollivier-Ricci curvature with the set of probability measures $\{\mu_i: i \in \mathcal{V}\}$ is defined along the geodesic path $[i, j]$ as

$$\kappa([i, j]) = 1 - \frac{W_1(\mu_i, \mu_j)}{d(i, j)}, \tag{7}$$

where $W_1(\mu_i, \mu_j)$ is the first Wasserstein distance between the probability measures μ_i and μ_j defined on $\mathcal{N}_i$ and $\mathcal{N}_j$, respectively, and is defined as

$$W_1(\mu_i, \mu_j) = \inf \sum_{k, \ell \in \mathcal{N}_i \times \mathcal{N}_j} d(k, \ell) \xi^{ij}(k, \ell). \tag{8}$$

The infimum is extended over all "coupling" measure $\xi^{ij}(k, \ell)$ defined on $\mathcal{N}_i \times \mathcal{N}_j$ and projecting on the first(second) factor as $\mu_i(\mu_j)$. More intuitively, $\xi^{ij}(k, \ell)$ is called *transference plan*. It tells us how much of the mass of k is transferred to ℓ, but it does not tell us anything about the actual path that the mass has to follow. $d(i, j)$ is the usual (distance) metric emanating from the edge weight ρ. The first Wasserstein distance W_1 is also referred to as the Earth Mover's Distance (EMD) in computer science applications.

Exact computation of the Ollivier-Ricci curvature can be computed via calculation of the Wasserstein distance using linear programming[6] and parallel computation[29,30].

References

1. Adamic, L. A. & Glance, N. The political blogosphere and the 2004 U.S. election. In *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, 36–43, https://doi.org/10.1145/1134271.1134277 (ACM Press, New York, New York, USA, 2005).
2. Udrescu, L. *et al*. Clustering drug-drug interaction networks with energy model layouts: Community analysis and drug repurposing. *Sci. Reports* **6**, 1–10, https://doi.org/10.1038/srep32745 (2016).
3. Zachary, W. W. An Information Flow Model for Conflict and Fission in Small Groups. *J. Anthropol. Res.* **33**, 452–473, https://doi.org/10.1086/jar.33.4.3629752 NIHMS150003 (1977).
4. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826, https://doi.org/10.1073/pnas.122653799 0112110 (2002).
5. Ollivier, Y. Ricci curvature of Markov chains on metric spaces. *J. Funct. Analysis* **256**, 810–864 0701886v4 (2009).
6. Wang, C., Jonckheere, E. & Banirazi, R. Wireless network capacity versus Ollivier-Ricci curvature under Heat-Diffusion (HD) protocol. *Proc. Am. Control. Conf.* 3536–3541, https://doi.org/10.1109/ACC.2014.6858912 (2014).
7. Wang, C., Jonckheere, E. & Banirazi, R. Interference constrained network control based on curvature. *Proc. Am. Control. Conf.* 2016-July, 6036–6041, https://doi.org/10.1109/ACC.2016.7526617 (2016).
8. Wang, C., Jonckheere, E. & Brun, T. Ollivier-Ricci curvature and fast approximation to tree-width in embeddability of QUBO problems. *Proc. 6th Int. Symp. on Commun. Control. Signal Process. (ISCCSP)* 2563–2566 (2014).
9. Wang, C., Jonckheere, E. & Brun, T. Differential geometric treewidth estimation in adiabatic quantum computation. *Quantum Inf. Process.* **15**, 3951–3966, https://doi.org/10.1007/s11128-016-1394-9 (2016).
10. Sandhu, R., Georgiou, T., Reznik, E., Zhu, L. & Kolesov, I. Graph Curvature for Differentiating Cancer. *Networks. Sci. Reports* **5**, 1–13, https://doi.org/10.1038/srep12323 (2015).
11. Sandhu, R. S., Georgiou, T. T. & Tannenbaum, A. R. Ricci curvature: An economic indicator for market fragility and systemic risk. *Sci. Adv.* **2**, 21–23 (2016).
12. Jost, J. & Liu, S. Ollivier's Ricci Curvature, Local Clustering and Curvature-Dimension Inequalities on Graphs. *Discret. Comput. Geom.* **51**, 300–322, https://doi.org/10.1007/s00454-013-9558-1 1103.4037 (2014).
13. Ariaei, F., Lou, M., Jonckheere, E., Krishnamachari, B. & Zuniga, M. Curvature of Indoor Sensor. *Network: Clustering Coefficient. EURASIP J. on Wirel. Commun. Netw.* **2008**, 213185, https://doi.org/10.1155/2008/213185 (2008).
14. Abbe, E. Community Detection and Stochastic Block Models: Recent Developments. *J. Mach. Learn. Res.* **18**, 1–86 (2018).
15. Breiger, R. L. & Pattison, P. E. Cumulated social roles: The duality of persons and their algebras. *Soc. Networks* **8**, 215–256, https://doi.org/10.1016/0378-8733(86)90006-7 (1986).
16. Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **74**, https://doi.org/10.1103/PhysRevE.74.036104 0605087 (2006).
17. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **84**, 1–19, https://doi.org/10.1103/PhysRevE.84.066106 (2011).
18. Wishart, D. S. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–D672, https://doi.org/10.1093/nar/gkj067 arXiv:1011.1669v3 (2006).
19. Fortunato, S. Community detection in graphs. *Phys. Reports* **486**, 75–174, https://doi.org/10.1016/j.physrep.2009.11.002 0906.0612 (2010).
20. Borgwardt, K. H. *The Simplex Method—A Probabilistic Analysis*, vol. 1 of *Algorithms and Combinatorics* (Springer-Verlag, New York, 1988).
21. Shirdhonkar, S. & Jacobs, D. W. Approximate earth mover's distance in linear time. *26th IEEE Conf. on Comput. Vis. Pattern Recognition, CVPR*, https://doi.org/10.1109/CVPR.2008.4587662 (2008).
22. Orlin, J. B. A polynomial time primal network simplex algorithm for minimum cost flows. *Math. Program.* **78**, 109–129, https://doi.org/10.1007/BF02614365 (1997).
23. Bauer, F., Jost, J. & Liu, S. Ollivier-Ricci curvature and the spectrum of the normalized graph Laplace operator. 1–20 1105.3803 (2011).
24. Jonckheere, E., Lou, M., Bonahon, F. & Baryshnikov, Y. Euclidean versus hyperbolic congestion in idealized versus experimental networks. *Internet Math.* **7**, 1–27, https://doi.org/10.1080/15427951.2010.554320 arXiv:0911.2538v1 (2011).
25. Banirazi, R., Jonckheere, E. & Krishnamachari, B. Heat diffusion algorithm for resource allocation and routing in multihop wireless networks. *GLOBECOM - IEEE Glob. Telecommun. Conf.* 5693–5698, https://doi.org/10.1109/GLOCOM.2012.6504028 (2012).
26. Banirazi, R., Jonckheere, E. & Krishnamachari, B. Dirichlet's principle on multiclass multihop wireless networks: Minimum cost routing subject to stability. *MSWiM 2014 - Proc. 17th ACM Int. Conf. on Model. Analysis Simul. Wirel. Mob. Syst.* 31–40, https://doi.org/10.1145/2641798.2641808 (2014).
27. Banirazi, R., Jonckheere, E. & Krishnamachari, B. Heat-Diffusion: Pareto optimal dynamic routing for time-varying wireless networks. *Proc. - IEEE INFOCOM* 325–333, https://doi.org/10.1109/INFOCOM.2014.6847954 (2014).
28. Banirazi, R., Jonckheere, E. & Krishnamachari, B. Minimum delay in class of throughput-optimal control policies on wireless networks. *Proc. Am. Control. Conf.* 2668–2675, https://doi.org/10.1109/ACC.2014.6859447 (2014).
29. Li, W., Ryu, E. K., Osher, S., Yin, W. & Gangbo, W. A Parallel Method for Earth Mover's Distance. *J. Sci. Comput.*, https://doi.org/10.1007/s10915-017-0529-1 (2017).
30. Ni, C.-C., Lin, Y.-Y., Gao, J., Gu, X. D. & Saucan, E. Ricci Curvature of the Internet Topology. *2015 IEEE Conf. on Comput. Commun. (INFOCOM)* **26**, 2758–2766, https://doi.org/10.1109/INFOCOM.2015.7218668 1501.04138 (2015).
31. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An open source software for exploring and manipulating networks. *Int. AAAI Conf. on Weblogs Soc. Media* 361–362, https://doi.org/10.1111/j.1939-1676.2011.0728.x (2009).

Acknowledgements

The authors gratefully acknowledge the support by the U.S. Army Research Office (ARO) under Grant No. W911NF-17-1-0076, the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award under Grant No. N66001-17-1-4044 support, the National Science Foundation Career award under Grant No. CPS/CNS-1453860, and the National Science Foundation (NSF) Grant CCF-1423624. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied by the Defense Advanced Research Projects Agency, the Department of Defense or the National Science Foundation.

Author Contributions

J.S. and P.B. conceived the research study, E.J. did the coarse geometry analysis and formulated the graph Ollivier-Ricci curvature approach, J.S. wrote the code and conducted the experiments, J.S., E.J. and P.B. analyzed the results. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-46079-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.