Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Detecting the Coarse Geometry of Networks

(revised version: November 2018)

by

Melanie Weber, Jürgen Jost, and Emil Saucan

Preprint no.: 97 2018

Detecting the Coarse Geometry of Networks

Melanie Weber
Princeton University
mw25@math.princeton.edu

Jürgen Jost
MPI MIS
jost@mis.mpg.de

Emil Saucan
ORT Braude & Technion
semil@braude.ac.il

Abstract

Clustering and sampling are key methods for the study of relational data. Learning efficient representations of such data relies on the identification of major geometric and topological features and therefore a characterization of its *coarse geometry*. Here, we introduce an efficient sampling method for identifying crucial structural features using a discrete notion of Ricci curvature. The introduced approach gives rise to a complexity reduction tools that allows for reducing large relational structures (e.g., networks) to a concise core structure on which to focus further, computationally expensive analysis and hypothesis testing.

1 Introduction

The identification of major geometric properties in relational data is key to efficient methods representation learning. Different notions of coarse geometry have been considered for analyzing structural features of relational data, mostly in the context of higher order (mesoscale) network structures and community detection [15, 8]. Informally, *coarse geometry* denotes the study of the geometric (or sometimes topological) properties, without considering fine-grained, small-scale features [9, 16, 10]. In network representations of relational data, this concept is closely related to the notion of a *network backbone*, that captures essential structural properties, such as clusters or communities and the "long-range" connections between distinct network regions. Another, well-studied application of coarse geometry is *sampling*: A "good" sample is representative of the crucial features of the full data set, i.e. it resembles its core structure and coarse geometric features. In the present paper, we connect these ideas to a curvature-based analysis of relational data. We will see below, that high curvature can be linked to high structural importance. It is then a natural idea to sample the nodes or edges with high curvature to identify the coarse geometry of the network. In fact, similar clustering approaches based on the combinatorial version of Gaussian curvature (*clustering coefficient*), have been successfully implemented in many data science applications (see [18] and references therein). One can extend the idea of curvature-based sampling to other notions of metric curvatures, such as Ricci curvature. The choice of Ricci curvature over other notions is grounded in the understanding that relational data is determined not by its members, but by the *relations* between them, suggesting an *edge-based* approach. Discrete Ricci curvature is defined as a function on the network's edges, allowing us to introduce the concept of *edge-based sampling*, as opposed to the mainly node-based approaches studied so far. Here, we will use Forman-Ricci curvature [7, 21, 22, 19], a simple and scalable discrete Ricci curvature that allows for incorporating given features of both nodes and edges as geometric information by defining suitable weights.

2 Theory

2.1 Complex Networks as Metric Measure Spaces

A natural idea for representing networks is to incorporate the node and edge weights into one expressive metric, thus rendering any weighted network into a metric space, whose geometric

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

properties can then be investigated with classical tools. A comprehensive approach was recently proposed in [6], the so-called *degree path metric*:

Definition 2.1 (Path degree metric) Let (G, w_v, w_e) be a weighted graph, where w_v denote the node weights, and w_e the edge weights. Then the function $\rho : V(G) \times V(G) \to [0, \infty)$, defined on the nodes $V(G)$ of G,

$$\rho(x, y) = \inf_{\pi=\{x_i\}} \sum_{i=1}^{n} \left(\max\{d(x_{i-1}), d(x_i)\}\right)^{-1/2} , \qquad d(x) = \frac{1}{w_v(x)} \sum_{y \sim x} w_e(x, y) ; \qquad (2.1)$$

represents a metric on G. Here, the infimum is taken over all paths $\pi = x = x_0 x_1 \dots x_n = y$, and d denotes the *weighted degree*.

This (global) metric allows for a notion of sampling with respect to the intrinsic geometry of graph, generalizing combinatorial notions such as the clustering coefficient.

Many of the geometric characteristics analyzed in continuous spaces have analogs in discrete regimes. Associations between these analogs can be used to derive discrete notions of curvature [3], notably the Olliver-Ricci curvature [13, 14] and the Forman-Ricci (short: *FR*) curvature [21, 22]. FR curvature was found to be especially useful for network analysis since its intuitive notion allows for efficient computation ($\sim O(|V(G)| + |E(G)|)$) that scales to large networks sizes. In its most general form, FR curvature is defined on CW cell complexes. Network representations of relational data $G = \{V(G), E(G)\}$ form regular, 1-dimensional cell complexes, in which case the following curvature function can be defined [21]:

$$\mathrm{Ric}_F(e) = w(e) \left(\frac{w(v_1)}{w(e)} + \frac{w(v_2)}{w(e)} - \sum_{\substack{e_{v_1} \sim e \\ e_{v_2} \sim e}} \left[\frac{w(v_1)}{\sqrt{w(e) w(e_{v_1})}} + \frac{w(v_2)}{\sqrt{w(e) w(e_{v_2})}} \right] \right)$$

The function is defined on each edge $e = (v_1, v_2) \in E(G)$ of the network, connecting vertices $v_1, v_2 \in V(G)$; w denotes the weights of edges and vertices.

2.2 Ricci Curvature-based Sampling

The key idea of curvature-based sampling is to choose sampling points whose metric density is inversely proportional to curvature (see Alg. 1(ii)). Most approaches in the literature are based on extrinsic curvature [18] thus requiring to find an isometric embedding in $\mathbb{R}^n$, a problem that is highly nontrivial for abstract (data) manifolds. Therefore, it is desirable to find a sampling method based on *intrinsic* curvature notions, such as Ricci curvature.

Such approaches are motivated by the close connection between volume growth rates and Ricci curvature [11]. While a detailed discussion of the underlying mathematical arguments is beyond the scope of this short paper, let us give the basic idea: The construction is based on so-called ε-nets (see Appendix A.2), that define efficient packings on manifolds. In this framework, the close connection between volume growth rates of the ε-balls and Ricci curvature can be studied explicitly. When generalizing to metric measure spaces, volumes are replaced by more general measures and the classical Ricci curvature by the generalized Rici curvature developed by Lott-Villani-Sturm [11, 20]. Observe that the graphs (ε-nets) rendered in the construction are coarsely equivalent (isometric) to the original metric measure space (see [17], Thm. 5.6, 5.11 and corollaries). Importantly, the obtained graph representation encodes the essential topology (homotopy) of the sampled space. The coarse reconstruction of the space is independent of the specific geometry of the manifold and the measure: It depends only on bounds on dimension, volume, curvature, and diameter. Motivated by these approaches, we will introduce a sampling method for networks, based on a discrete notion of Ricci curvature, to detect the coarse geometry.

A cornerstone of representation learning is the embedding, with low distortion, of a given data set into a model space with respect to its relational structure. Classic examples are Euclidean embeddings, where the ambient space is usually $\mathbb{R}^n$, for some n large enough. More recently, embeddings of networks into hyperbolic space $\mathbb{H}^d$ have been studied [12, 5]. Here, we ask, whether one can construct a *coarse embedding* of a weighted graph, viewed as a metric measure space. By a coarse embedding of a metric space (X, d) into another, we mean a map $i : X \to Y$, such that there exist increasing, unbounded functions $\eta_1, \eta_2 : \mathbb{R} \to \mathbb{R}$, such that $\eta_1(d(x_1, x_2)) \leq d(i(x_1), i(x_2)) \leq \eta_2(d(x_1, x_2))$, for any $x_1, x_2 \in X$. Recall the following:

Definition 2.2 Given a set X, a *kernel* is a symmetric function $k : X \times X \to \mathbb{R}$, i.e. $k(x, y) = k(y, x)$, for any $x, y \in X$. A kernel k is said to have

1. *positive type*, if the matrix $K_m = \{k(x_i, x_j)\}_{i,j=1}^m$ is positive semidefinite for all $m \in \mathbb{N}$;
2. *negative type*, if the matrix $K_m = \{k(x_i, x_j)\}_{i,j=1}^m$ is negative semidefinite for all $m \in \mathbb{N}$.

The curvature operator Ric_F (see, e.g., [3]) is symmetric in the nodes u, v of an edge $e = (u, v)$, i.e., defines a kernel k_F. Recall, that for positive semidefinite operators (such as Laplacians), the corresponding kernels are also positive semidefinite. By a direct application of classical results (see, e.g. [16], Thm. 11.15a), there exists a map $\varphi : X \to \mathcal{H}$, where $\mathcal{H}$ is a real Hilbert space, such that $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$, for all $x, y \in X$. However, Forman-Ricci curvature is generally everywhere non-positive. In fact, $\mathrm{Ric}_F(e = (u, v))$ is *not* negative only if both u, v have at most degree 2 (i.e., only in n-cycles and in small degenerate structures). Therefore, we can assume that for non-degenerate graphs $\mathrm{Ric}_F(e) < 0$ everywhere. Thus, a mapping into a real Hilbert space, as for the Laplacians, is not possible. Instead, we relate the kernel k_F to a (positive) notion by setting $k_F^* = e^{-k_F}$[1] and map to a Hilbert space as described above. Therefore, the existence of a *coarse* embedding of a graph into a (real) Hilbert space follows from (i) e^{-k_F} being a positive kernel and (ii) the edges $\{e = (u, v) \,|\, k(e) < k < 0\}$ with curvature bounded above generating the *coarse structure* of the network.

3 Methods

Algorithm 1 Curvature-based sampling

Input: $G = \{V, E, w_v, w_e\}$; (i) k, (ii) r
for $u, v \in V,\ u \sim v$ **do**
    $k_F(u, v) \leftarrow Ric_F(e = (u, v))$
end for
(i) $S \leftarrow \{e = (u, v) \,|\, k(e) < k < 0\}$
(ii) $k_F^* \leftarrow e^{-k_F},\ \hat{k}_F^* \leftarrow reconstruct(kernelPCA(k_F^*), r);\ S \leftarrow \{e = (u, v) \,|\, \hat{k}_F^*(e) \neq 0\}$
Output: $G' = \{V\big|_S, E\big|_S, w_v\big|_S, w_e\big|_S\}$

The characterization of substructures, such as motifs or communities, has been a cornerstone of network analysis since its early days [8]. More recently, higher-order structures, such as the notion of the *network backbone*, have gained major interest [4]. The key idea is the reduction of a system to its most essential elements and the relations between them. In the case of large-scale networks, the backbone comes with the promise of complexity reduction that could make complex structures accessible to computational tools that are otherwise too costly. We consider the following (informal) notion:

Definition 3.1 (Network backbone) We denote the *backbone* of a network $G = \{V, E\}$ as a sub-network $G' = \{V', E'\}$ ($V' \subseteq V$, $E' \subseteq E$) that captures structurally important nodes (*hubs*) and edges (*bridges*). A node is typically termed *hub* if it has a high degree and a high betweenness centrality. *Bridges* denote edges that govern the mesoscale structure of G, for instance by forming long-range connections between communities. The backbone G' is *structure-preserving*, i.e., its structural features (e.g., node degree distribution, community structure) are representative of G.

We propose the identification of the network backbone through *curvature-based sampling*: Approach Alg. 1(i) is based on the fact that high absolute curvature is strongly related to the structural importance of an edge. A combinatorial reasoning behind this observation is apparent from the rewritten curvature function: $\mathrm{Ric}_F(e) = w(v_1) + w(v_2) - \sum_{\substack{e_{v_1} \sim e \\ e_{v_2} \sim e}} \left[w(v_1) \sqrt{\frac{w(e)}{w(e_{v_1})}} + w(v_2) \sqrt{\frac{w(e)}{w(e_{v_2})}} \right]$. We see that high absolute curvature results from high node degrees in vertices v_1 and v_2 and from the edge weight $w(e)$ being large compared to those of parallel edges ($w(e_{v_1})$, $w(e_{v_2})$). Since high-degree nodes (hubs) form the centers of the major network communities, the edges that form strong connections between them bridge the corresponding communities [2] – forming the network backbone. As

[1] In fact, for every $t > 0$, the kernel $k_F^{*,t}(x, y) = e^{-t k_F(x,y)}$ is of positive type (see [16], Prop. 11.12.)

discussed in the theory section, we can sample those crucial structural features by computing Forman-Ricci curvature across the network and selecting the edges with the highest absolute curvature. This allows us to empirically determine the network backbone in a computationally efficient way. Approach Alg. 1(ii) implements the kernel-based approach discussed in section 2, i.e., choosing sampling points whose metric density is inverse proportional to curvature.

4 Experiments

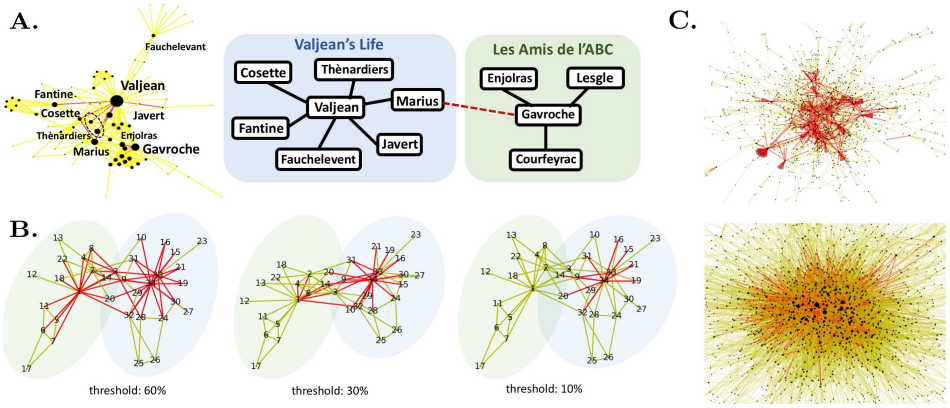

Figure 1: **A:** Network backbone (marked red) for *Les Miserables* identifies major storyline. **B:** Influence of backbone threshold for sampling *Zachary's Karate Club*. **C:** Backbone sampling in large networks. All data sets can be found in ICON [1].

We perform curvature-based sampling on a weighted network of character co-occurrences in Victor Hugo's *Les Miserables*. The backbone with threshold 5% (i.e., sampling bound k is chosen such that it is larger than 5% of the curvature values) is marked in red (Fig 4A). The sampling identifies relationships between Valjean and other major characters that are central to Valjean's storyline, as well as a second cluster of revolutionaries around Gavroche. Both clusters merge ("bridge") when setting the backbone threshold to 11%. This indicates that we can identify major relations in the data and high-level structural information by sampling for edges with high absolute curvature. In a second experiment, we compare the known community structure of *Zachary's karate club* with the sampled backbone (Fig. 4B, shown in red). We observe, that keeping the top 30% edges highest absolute curvature, covers the major structural features: The instructor (node 1) and the president (node 34), as well as their neighborhoods, are in the backbone, forming the two known clusters. Note that the sampled backbone also includes the bridges between the two clusters (i.e., the joint neighbors of 1 and 34). In a third experiment (Fig. 4C) we identify the backbone in two larger data sets (yeast transcriptome (top) and bible wordnet (bottom) demonstrating a potential application of the method as complexity reduction tool.

5 Discussion

We introduced a curvature-based sampling method for identifying structurally important structural features in relational data. We sample a "backbone" with respect to Forman-Ricci curvature, an efficiently computable discrete notion of Ricci curvature on networks. The identification of this core structure is of great importance for the structural analysis of large networks: It allows us to reduce large networks to its core structure on which computationally expensive network hypothesis testing and further network analysis become feasible. Ongoing work includes a statistical analysis of the preservation of coarse geometric features under curvature-based sampling and a comparison with related sampling methods (sampling accuracy, computational efficiency) on large-scale real-world and synthetic relational data.

References

[1] Ellen Tucker Aaron Clauset and Matthias Sainz. The Colorado Index of Complex Networks. `https://icon.colorado.edu/`, 2016.

[2] L. Backstrom and J. Kleinberg. Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 831–841, 2014.

[3] F. Bauer, B. Hua, J. Jost, S. Liu, and G. Wang. *Modern Approaches to Discrete Curvature. Lecture Notes in Mathematics*, volume 2184, chapter The Geometric Meaning of Curvature: Local and Nonlocal Aspects of Ricci Curvature. Springer, Cham, 2017.

[4] Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

[5] Christopher De Sa, Albert Gu, Christopher Re, and Frederic Sala. Representation tradeoffs for hyperbolic embeddings. *arXiv preprint arXiv:1804.03329*, 2018.

[6] Jozef Dodziuk and Wilfrid Kendall. Combinatorial laplacians and isoperimetric inequality. *Res. Notes Math. Ser.*, 150:68–74, 01 1986.

[7] R. Forman. Bochner's Method for Cell Complexes and Combinatorial Ricci Curvature. *Discrete and Computational Geometry*, 29(3):323–374, 2003.

[8] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1 – 44, 2016. Community detection in networks: A user guide.

[9] M. Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. Birkhauser, Boston, 1999.

[10] Masahiko Kanai. Rough isometries, and combinatorial approximations of geometries of non-compact riemannian manifolds. *J. Math. Soc. Japan*, 37(3):391–413, 07 1985.

[11] John Lott and Cedric Villani. Ricci curvature for metric-measure spaces via optimal transport. *Annals of Mathematics*, 169(3):903–991, 2009.

[12] Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *Proceedings of the Thirty-fifth International Conference on Machine Learning (to appear)*, 2018.

[13] Y. Ollivier. Ricci curvature of Markov chains on metric spaces. *Journal of Functional Analysis*, 256(3):810–864, 2009.

[14] Y. Ollivier. A survey of Ricci curvature for metric spaces and Markov chains. *Probabilistic approach to geometry*, 57:343–381, 2010.

[15] Tiago P. Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Phys. Rev. E*, 92:042807, Oct 2015.

[16] J. Roe. *Lectures on Coarse Geometry*. AMS, Providence, RI, 2003.

[17] E. Saucan. Curvature based triangulation of metric measure spaces. *Contemporary Mathematics*, 554:207–227, 2011.

[18] Emil Saucan, Eli Appleboim, and Yehoshua Y. Zeevi. Sampling and reconstruction of surfaces and higher dimensional manifolds. *Journal of Mathematical Imaging and Vision*, 30(1):105–123, Jan 2008.

[19] R.P. Sreejith, K. Mohanraj, J. Jost, E. Saucan, and A. Samal. Forman curvature for complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, (6), 2016.

[20] Karl-Theodor Sturm. On the geometry of metric measure spaces. *Acta Math.*, 196(1):65–131, 2006.

[21] Melanie Weber, Emil Saucan, and Jürgen Jost. Characterizing complex networks with forman-ricci curvature and associated geometric flows. *Journal of Complex Networks*, 5(4):527–550, 2017.

[22] Melanie Weber, Emil Saucan, and Jürgen Jost. Coarse geometry of evolving networks. *Journal of Complex Networks*, 2017.

A Mathematical Context

A.1 Motivation for path degree metric

This path degree metric introduced in the main text has the benefit of being both efficiently computable (and therefore scalable) and expressive: It is closely related to the random walk on a graph, given by the jump from a vertex x to an adjacent vertex (i.e. neighbor) vertex y, with probability

$$p = w_e(x,y)\left(\sum_{y'\sim x} w_e(x,y')\right)^{-1}.$$

Since, in addition, the probability of not leaving x at the time t is $e^{-d(x)}$, it follows that the larger the degree, the faster the random walk departs from x. In addition, from the definition of $\rho(x,y)$ it follows that the larger the degree of x or y, the closer the two vertices are to each other, i.e., the faster the jump along an edge, the shorter the edge is. Note that, given that the degrees of x and y might be unequal, the jumping time from x to y is not necessarily symmetric, instead, ρ prefers the larger degree with faster jumping time.

A.2 Efficient packings on manifolds

Consider the following notion of *efficient packings*:

Definition A.1 *Let $p_1,\dots,p_{n_0}$ be points $\in M^n$, satisfying the following conditions:*

1. *The set $\{p_1,\dots,p_{n_0}\}$ is an ε-net on M^n, i.e. the balls $\beta^n(p_k,\varepsilon)$, $k=1,\dots,n_0$ cover M^n;*
2. *The balls (in the intrinsic metric of M^n) $\beta^n(p_k,\varepsilon/2)$ are pairwise disjoint.*

Then the set $\{p_1,\dots,p_{n_0}\}$ is called a minimal ε-net and the packing with the balls $\beta^n(p_k,\varepsilon/2)$, $k=1,\dots,n_0$, is called an efficient packing. The set $\{(k,l)\,|\,k,l=1,\dots,n_0 \text{ and } \beta^n(p_k,\varepsilon)\cap\beta^n(p_l,\varepsilon)\neq\emptyset\}$ is called the intersection pattern of the minimal ε-net (of the efficient packing).

This notion encodes the close connection between volume growth rates in manifolds and Ricci curvature. It is, therefore, only natural to seek the generalization of this construction to metric measure manifolds, where volumes are replaced by more general measures, and instead of the classical Ricci curvature one employs the generalized Rici curvature developed by Lott-Villani [11] and Sturm [20]. Such an extension of the classical case construction to the metric measure spaces context does, indeed exists [17]. More precisely, one gas the following theorem:

Theorem A.2 *Let (M_1^n,d_1,ν_1), (M_2^n,d_2,ν_2), $\nu_i=e^{-V_i}d\text{Vol}$, $V_i\in\mathcal{C}^2(\mathbb{R})$, $i=1,2$ be smooth, compact metric measure spaces satisfying $\text{CD}(K,N)$ for some $K\in\mathbb{R}$ and $1<N<\infty$, and such that $\text{diam}M_i^n<D$, $\text{Vol}M_i^n<v$, $i=1,2$ and, moreover, having the same lower bound k on their sectional curvatures. Then there exists $\varepsilon=\varepsilon(N,K,k,D,v)$ such that, if M_1^n,M_2^n have minimal packings with identical intersection patterns, they are homotopy equivalent.*

Here $\text{CD}(K,N)$ denotes the generalized Ricci curvature of Lott, Villani, and Sturm.