Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Edge-based analysis of networks:
Curvatures of graphs and hypergraphs

by

Marzieh Eidi, Amirhossein Farzam, Wilmer Leal,
Areejit Samal, and Jürgen Jost

Edge-based analysis of networks: Curvatures of graphs and hypergraphs

Marzieh Eidi[1], Amirhossein Farzam[1], Wilmer Leal[1,2], Areejit Samal[3], and Jürgen Jost[1,4]

[1]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany
[2]Bioinformatics Group, Department of Computer Science, Universität Leipzig, 04107 Leipzig, Germany
[3]The Institute of Mathematical Sciences (IMSc), Homi Bhabha National Institute (HBNI), Chennai 600113, India
[4]The Santa Fe Institute, Santa Fe, New Mexico 87501, USA

October 29, 2020

Abstract

The relations, rather than the elements, constitute the structure of networks. We therefore develop a systematic approach to the analysis of networks, modelled as graphs or hypergraphs, that is based on structural properties of (hyper)edges, instead of vertices. For that purpose, we utilize so-called network curvatures. These curvatures quantify the local structural properties of (hyper)edges, that is, how, and how well, they are connected to others. In the case of directed networks, they assess the input they receive and the output they produce, and relations between them. With those tools, we can investigate biological networks. As examples, we apply our methods here to protein-protein interaction, transcriptional regulatory and metabolic networks.

1 Introduction

A central paradigm of structuralism [1, 2] is the analysis of structural relations regardless of the identity of the elements involved. That is, a structure is conceived in terms of the *relations* between elements. One wants to understand the types of relations, rather than the nature of the elements. This paradigm is obviously also fundamental for the analysis of empirical networks, be they from the biological sciences or other domains. Such an analysis then again abstracts from the specific content of the elements and concentrates on the formal relations between them. In that manner, one can both find universal features that

1

hold across a wide range of networks from different domains, and properties that are specific to particular empirical domains.

For that purpose, many different measures have been developed. Some of these measures, like for instance the assortativity (see for instance [3, 4]), are of a global nature, that is, associate some number to the entire network. Such a number is usually an average or perhaps, like the diameter of a network, an extremum of locally measured quantities. In any case, the basis for such global measures is to first develop local measures. For a more refined analysis, one can then also look at the statistics of those local measures, instead of lumping them together in a single number (for instance [5, 6] for assortativity).

Some of these local measures require global computations in the network; for instance, for computing the diameter, one needs to evaluate the distances between any two elements. Therefore, some of these measures are difficult to evaluate in practice for networks of more than a moderate size. Others, including those that we shall concentrate on in this contribution, require only local computations and can be very easy to evaluate.

Now, somewhat surprisingly in view of the above structuralist paradigm, many of the local measures assign numbers to the elements of the network, rather than directly to its relations. The most basic one here is the degree of an element, the number of relations that it participates in. More global measures for instance evaluate the robustness of the network in terms of how many or which elements need to be eliminated in order to disconnect the network. See for instance [7, 8].

In this situation, we and our collaborators have developed the research paradigm of a relation based analysis of networks (for instance [9, 10, 11, 12, 13, 14, 15, 16, 17, 6]. That is, we evaluate relations and associate measures to them whose statistics across the network then can provide structural insight.

There is another shortcoming of much of traditional network analysis. It tries to represent all structures as graphs, that is, considers only pairwise relations. For instance, a relation between three elements is simply broken up into three pairwise relations. That may, however, suppress some important structural insight. Take the example of scientific collaborations. From preprint repositories in the internet, it is easy to extract patterns of collaborations from coauthorships between authors. There are some single author papers, but of more interest are those written by several authors. There may be more than two authors involved in some paper, say A, B, and C. Of course, one could reduce it to pairwise relations and say that any two of them are coauthors. But there may be more structure. For instance, there may also exist a two-author paper by A and B, no such paper between A and C, and a paper of B and C with two other authors D and E. This is obviously not captured by the pairwise relations, and for a more adequate model of the structure of scientific collaborations, we should rather use a hypergraph instead of a simple graph. In a hypergraph, a hyperedge can connect any number of elements. See for instance [18, 19, 20, 21, 22]. In computer science, directed hypergraphs are also known as Petri nets [23, 24]. They were originally proposed by Petri as models of chemical reactions. Over the years, while not as widely employed as graphs, they have found applications

in many fields, for instance recently as models of coupled dynamics in statistical physics [25, 26, 27], of social contagion [28] and for knowledge representation in natural language processing [29].

In this contribution, we shall summarize relation-based measures both for graphs, that being the simplest case, and for hypergraphs.

2 The idea of curvatures

The name *curvature* derives from its origin in differential geometry. Originally, curvature was an infinitesimal quantity, obtained by taking second derivatives of functions describing shapes of smooth objects, like curves or surfaces. In Riemannian geometry, curvatures obtained a deeper conceptual significance, as tensors encoding the geometric invariants of Riemannian metrics of smooth manifolds [30]. In particular, the Ricci tensor is fundamental not only in Einstein's theory of general relativity and in elementary particle physics (the Calabi-Yau manifolds of string theory, for instance, are characterized by the vanishing of the Ricci tensor), but it also permeates much of modern research in Riemannian geometry. While Ricci curvature in Riemannian geometry again is computed infinitesimally, by taking second derivatives of the metric tensor, it essentially encodes local property, like the average divergence of geodesics or the growth of the volume of balls as a function of their radii. Moreover, Bochner type identities link it to other important geometric quantities, like the first eigenvalue of the Laplace operator. See for instance [31] for a survey.

Since such objects and properties are also meaningful and important in metric spaces that are more general than Riemannian manifolds, alternative definitions of Ricci curvatures have been proposed that are formulated in terms of local quantities and no longer depend on taking derivatives. Several of these definitions turned to be also meaning- and useful for graphs, and we have extended them to hypergraphs and are exploring their properties. Here, we shall not recount the history in detail, but rather systematically develop a conceptual approach that is in line with the paradigm of structuralism described at the beginning. We only note the curious fact that these concepts, although extremely natural from a structuralist perspective, were not developed directly, but inspired by concepts in a different, and more highly developed branch of mathematics, Riemannian geometry.

3 How relations connect

Abstractly, there are different types of relations. They can vary with respect to the number of elements involved, they can be symmetric or directed, that is, distinguish between inputs and outputs, and they may also carry weights. The simplest case are binary, symmetric and unweighted relations. Such a web of relations is then modelled by an undirected and unweighted graph whose vertices stand for the elements in question and whose edges represent the presence of a

relation between the two vertices they connect. For simplicity, we also assume that the graph is simple, that is, there is at most one edge between any two vertices, and that it is connected, that is, by passing from edge to edge we can reach any vertex from any other one, although these assumptions are not essential for any of the sequel. So, we start with that case.

We want to assess how a relation, that is, an edge of such a graph, sits in the web of relations, that is, how it relates to other relations. Two edges are called neighbors when they share a vertex. We can then already define the simplest concept, called *Forman-Ricci curvature*, because it was introduced by Forman [32] as an analogy with the Ricci curvature of Riemannian geometry (the analogy relates to the role it plays in Bochner type identities). We define the degree of an edge e as

$$\deg(e) := \#(\text{neighbors of } e), \tag{1}$$

and define its Forman-Ricci curvature as

$$F(e) := 2 - \deg(e). \tag{2}$$

The 2 and the minus sign are somewhat unfortunate for our purposes, but they are there because of the analogy with the well-established Ricci curvature of Riemannian geometry, and they are useful from an abstract geometric perspective.

When the edge e connects the vertices v, w, we can also assess their contribution to the number of neighbors of e. We let $\deg_v(e)$ be the number of edges that share with e the vertex v. Then, obviously,

$$F(e) = 2 - (\deg_v(e) + \deg_w(e)). \tag{3}$$

Instead of the sum of the degrees, we may also consider their difference. When the edge is not directed, there is no intrinsic structural difference between the two vertices that it connects, and so, it is natural to take the absolute value of the difference and define the *degree difference* [6] as

$$\daleth(e) := |\deg_v(e) - \deg_w(e)|. \tag{4}$$

Let us interpret the geometric significance of these quantities. $\daleth(e)$ is large when e connects vertices of different types, a well-connected one from which many further edges emanate, and a less well-connected one from which only fewer edges originate. The statistics of this quantity therefore quantify to what extent the network is assortative, that is, typically connect similar vertices (small $\daleth(e)$), or disassortative, that is, typically connect dissimilar vertices (large $\daleth(e)$). This is important, for instance, because social networks tend to be assortative [33] (well connected people like to link with other well connected people, and this further improves their position in social networks). In contrast, $F(e)$ is very negative, that is, has a particularly large absolute value when both ends of an edge are well connected. Such edges may play a very important role in the network. In

4

fact, we have found [13] that a quantity that needs a global computation, edge-betweenness centrality (see [7]), is statistically well correlated with $F(e)$. This edge-betweenness centrality measures how many shortest connections between pairs of vertices in the network pass through that particular edge. The computation of that quantity is expensive because all shortest connections between any two vertices have to be evaluated. In contrast, the computation of $F(e)$ is very quick and easy, because only local neighborhoods have to be evaluated.

Edges with large $|F(e)|$ also play an important role for spreading in the network because from its vertices many other vertices in the network can be reached in a single step. There is one issue here, however. Edges from the two vertices v, w of e may end at the same vertex z, that is, v, w, z may form a triangle. In that case, they would not contribute to spreading into different directions. Or the endpoint of an edge from v and that of an edge from w may be connected themselves by an edge. That is, they form a quadrangle together with v and w. Again, that does not really constitute spreading into different directions. It is possible to address this issue by inserting two-dimensional faces into such triangles and perhaps also into quadrangles, and then to evaluate the Forman curvature of the resulting simplicial or polyhedral complex. Those faces would then increase the Forman curvature and make it less negative or even positive. See for instance [14].

This aspect is taken care of in a different way by a more refined concept of Ricci curvature, the *Ollivier-Ricci* curvature introduced in [34]. For that purpose, consider the edge $e = (v, w)$ and let $e_v = (v, v_1)$ and $e_w = (w, w_1)$ be edges emanating from v and w, respectively. We then define their distance w.r.t. e as

$$d_e(e_v, e_w) := d(v_1, w_1) \tag{5}$$

where $d(v_1, w_1)$ denotes the distance between v_1 and w_1 in the network, that is, the minimal edges that have to be traversed for getting from v_1 to w_1. Let E_v be the set of edges that have v as a vertex, and let $|E_v|$ be its cardinality. We then define a probability measure μ_v on the set of all edges E by giving each edge $e_v \in E_v$ the weight $\frac{1}{|E_v|}$ and all edges not in E_v the weight 0. We then define the Ollivier-Ricci curvature [34] of the edge $e = (v, w)$ as

$$O(e) := 1 - W_1(\mu_v, \mu_w) \tag{6}$$

where W_1 is the 1-Wasserstein distance between μ_v and μ_w,

$$W_1(\mu_v, \mu_w) := \inf_{p \in \Pi(\mu_v, \mu_w)} \sum_{(e_1, e_2) \in E \times E} d_e(e_1, e_2) p(e_1, e_2) \tag{7}$$

and $\Pi(\mu_v, \mu_w)$ is the set of measures on $E \times E$ that project to μ_v and μ_w, resp. We thus try to arrange the two collections E_v, E_w of edges sharing one of their endpoints with e in an optimal manner, that is, that the average distances of the arranged pairs become as small as possible. We note that the sets E_v and E_w both include the edge $e = (v, w)$ that we are evaluating. This convention is only needed to let our definition agree with that originally proposed in the literature,

but could otherwise be abandoned, to make the definition more natural in the present context.

In order to evaluate (6), we have to optimize the arrangement between the edges in E_v and E_w, to make the transportation cost as small as possible. Since this is a quantity all edges in those two edge sets, it is not necessarily the case that an optimal transport plan arranges each edge e_1 in E_v with the edge e_0 in E_w closest to it. There might be some competition, as there might be other edges $e_2, e_3, \ldots$ for which e_0 is closest. But even if there is no such competition, it might be overall more beneficial to arrange e_1 with an edge different from e_0. Also, because of the normalization, the edges in E_v and E_w have fractional weights, and if the cardinalities of the two edge sets are different, also the corresponding weights are different, necessitating an arrangement where some part of an edge in E_v is arranged with some part of an edge in E_w, and other parts with other ones.

Notwithstanding these complications, let m_i be the fraction of edges in E_v that are moved a distance i in some optimal transport plan (such an optimal arrangement need not be unique, but that does not matter for our discussion). Then [35]

$$O(e) = m_0 - m_2 - 2m_3. \tag{8}$$

In particular, moving an edge a distance 1 does not contribute at all to $O(e)$. (While m_1 itself does not appear in (8), its computation is nevertheless needed as an intermediate step for computing m_2 and m_3.) Distance 0, that is, when e participates in a triangle, has a positive contribution. A pentagon, that is, distance 2, has a negative contribution, but not as a negative as the maximal distance, that can occur in a transportation plan, which is 3. This simple formula thus encodes the essential features of Ollivier-Ricci curvature. In fact, we could simply take (8) as the definition of $O(e)$, instead of utilizing the more complicated formula (7).

More generally, the Ollivier-Ricci curvature is related to the clustering coefficient, that is, the relative frequency of triangles in the network [36].

Protein-protein interaction networks

To illustrate an application of these structural measures to empirical data, we have studied the protein-protein interaction (PPI) networks in human [37], with 8275 nodes and 52569 edges, and fission yeast *S. pombe* [38], with 1306 nodes and 2278 edges. The edges in these network represent binary interactions between the pair of proteins represented as nodes. These undirected and unweighted networks are disconnected with several components, however, they both include a giant component. The giant component consists of 8152 nodes and 52036 edges in the human PPI network, and of 1306 nodes and 2278 edges in fission yeast PPI network. We have computed the Forman-Ricci curvature, Ollivier-Ricci curvature and degree difference of edges in these networks, and their distributions are shown in Figure 1.

In the human PPI network, while Ollivier-Ricci curvature has a unimodal

6

distribution, the bimodal distribution of Forman-Ricci curvature in Figure 1 signals an evident heterogeneity in the space of protein-protein interactions in the giant component; a major group of interactions are distributed around a relatively small-valued mode, and a small group of interactions between proteins that are, in average, involved in a signficantly larger number of interactions. The degree difference distribution indicates that, although the majority of interactions are between proteins with relatively similar degree, a noticeable proportion of the edges have a considerably large degree difference, which can be as large as 497. This observation is in line with the fact that this network is moderately disassortative with assortativity value ~ -0.119.

Unlike the Ollivier-Ricci curvature distribution of the human PPI network, the fission yeast PPI network has a trimodal distribution of the Ollivier-Ricci curvature, reaching its global mode at curvature value 0. In fact, in the PPI network for fission yeast, all three measures have multimodal distributions, as demonstrated in Figure 1. Interestingly, the peaks over highly negative values of Forman-Ricci curvature have larger frequencies than those over the moderately negative values. A similar phenomenon is observed in the degree difference distribution of the fission yeast PPI network. The global degree assortativity of the fission yeast PPI network is ~ -0.237. This means that the fission yeast PPI network is considerably more disassortative than the human one, which is explained by the more substantial proportion of interactions in fission yeast between proteins with significantly different degrees. Thus, we see that the distribution of curvature and degree difference values can point us to biologically relevant properties of the interaction statistics in the PPI networks of different species.

4   Directed graphs

It is not only the case that the preceding constructions extend to directed graphs, but in fact, they become even more natural in that context. Curvature concepts for directed graphs have been systematically developed and evaluated in [14] (see also see [12]). Here, we shall formulate the concepts in such a manner that they will naturally generalize to hypergraphs.

Thus, let $e = [v, w]$ be a directed edge with tail v and head w, that is, going from v to w. The input of e at its tail v then are all edges that have v as their head; let their number be $\deg_{in}(e)$. Similarly, $\deg_{out}(e)$ denotes the number of output edges of e, that is, those that have w as their tail. We may then put [39]

$$F_{\rightarrow}(e) := 2 - \deg_{in}(e) - \deg_{out}(e). \tag{9}$$

We could also form alternative expressions by considering the numbers of edges that have v as their tail and/or of those that have w as their head. Similarly for the next expression, the directed degree difference [6]

$$\daleth_{\rightarrow}(e) := \deg_{out}(e) - \deg_{in}(e). \tag{10}$$

7

$F_\rightarrow(e)$ now is very negative, or equivalently, $\deg_{in}(e) + \deg_{out}(e)$ is very large for those edges that receive a lot of input and produce a lot of output. $\daleth_\rightarrow(e)$ is positive for those edges that are productive in the sense that they produce more output than they receive input, or that lead to more diversification. It is negative for edges that are receptive, that is collect more input than emit as output.

Likewise, we can define the Ollivier-Ricci curvature $O_\rightarrow(e)$ of a directed edge [35] by computing the optimal transportation distance between its input and its output. When there are no shorter connections between inputs and outputs than those going through e itself, then $O_\rightarrow(e)$ assumes its smallest possible value -2. In contrast, when inputs coincide with outputs, that is, if there is a directed triangle from a vertex u to itself, where u produces both an input of e and receives an output of e, then this yields a positive contribution. In fact, formula (8) perfectly extends to the directed case.

Let us recall the procedure in detail. For the directed edge $e[v, w]$, we define two measures,

$$\begin{cases} \text{if } e \text{ has no incoming edges:} & \mu_{in}(e) = 1 \\ \text{if } e \text{ has } n_1 \text{ incoming edges:} & \mu_{in}(e_1) = \frac{1}{n_1} \text{ for each incoming edge} \end{cases} \quad (11)$$

$$\begin{cases} \text{if } e \text{ has no outgoing edges:} & \mu_{out}(e) = 1 \\ \text{if } e \text{ has } n_2 \text{ outgoing edges:} & \mu_{out}(e_2) = \frac{1}{n_2} \text{ for each outgoing edge} \end{cases} \quad (12)$$

and $\mu_{in}(e') = \mu_{out}(e') = 0$ for all edges e' not occurring in those formulae.

The first cases, that is, where there are no incoming edges at the tail or no outgoing edges at the head, that is, where the tail is a source or the head is a sink, represent complications that would not arise in the undirected case. As they are easily handled, we shall mostly ignore them. In any case, both measures are normalized to have total mass 1 and thus are probability measures. We then define the distance between an edge e_1 occurring in (11) and an edge e_2 occuring in (12) as

$$d_e(e_1, e_2) = \quad \text{minimal number of edges needed to get from the tail of } e_1$$
$$\text{to the head of } e_2. \quad (13)$$

We then put again

$$O_\rightarrow(e)_\rightarrow := 1 - W_1(\mu_{in}, \mu_{out}) \quad (14)$$

where W_1 now is the 1-Wasserstein distance between μ_{in} and μ_{out},

$$W_1(\mu_{in}, \mu_{out}) := \inf_{p \in \Pi(\mu_{in}, \mu_{out})} \sum_{(e_1, e_2) \in E \times E} d_e(e_1, e_2) p(e_1, e_2) \quad (15)$$

and $\Pi(\mu_{in}, \mu_{out})$ is the set of measures on $E \times E$ that project to μ_{in} and μ_{out}, respectively. E here is the set of directed edges of the directed graph under consideration. We again have the important formula [35]

$$O_\rightarrow(e) = m_0 - m_2 - 2m_3, \quad (16)$$

8

where m_i is the number of edges that have to be transported by the distance i in an optimal transport plan. Again, for two given edges in the in- and output of e, that distance might be larger than the distance (13).

Thus, the directed curvature notions evaluate flows through edges. As in [14], we may also evaluate flows through vertices by taking the difference between the sum of the Ricci curvatures of the incoming edges and that for the outgoing edges. Moreover, in [14], also notions of augmented Forman curvature were developed for directed networks. Augmentation means that one inserts two-dimensional faces into triangles of edges. Such triangles then increase the curvature, and thereby decrease the difference between Forman and Ollivier type curvatures. Here, however, we do not explore that direction.

Transcriptional regulatory networks

To illustrate an application to directed networks, we have studied the transcriptional regulatory network (TRN) of the important human pathogen *Mycobacterium tuberculosis* [40], with 2547 nodes and 6581 edges. The *M. tuberculosis* TRN was constructed based on ChIP-seq data for 143 transcription factors (TFs) [40]. In this directed and unweighted network, each directed edge signifies the regulatory control by a TF of a target gene. In other words, the source nodes in this directed network are TFs while target nodes are target genes. In Figure 2, we show the distribution of the Forman-Ricci curvature, Ollivier-Ricci curvature and degree difference of directed edges in the *M. tuberculosis* TRN. In Figure 2a, it is seen that the edges are densely concentrated around Forman-Ricci curvature value 0, and this indicates that the majority of the edges have a tail vertex with small in-degree and a head vertex with small out-degree. Likewise, most edges have zero or small value of directed degree difference, and this indicates that the in-degree of the tail vertex and the out-degree of the head vertex for most edges are rather similar. There are also 24 vertices with out-degree greater than 100, which can explain the long tail in both Forman-Ricci curvature and degree difference distributions in Figure 2. On the other hand, the Ollivier-Ricci curvature of the edges in this TRN has a multimodal distribution, with major peaks corresponding to curvature values 0, -1, -0.5, and -0.75.

5 Weighted graphs

The extension of all discussed concepts to weighted graphs is straightforward. One simply counts each edge with its weight. It is therefore not necessary to develop the details here.

6 Hypergraphs

The preceding concepts are set up in such a manner that they naturally extend to hypergraphs. We directly consider directed hypergraphs. A directed hyper-

graph may have several nodes through which it receives inputs and several nodes through which it produces outputs. An important example is a chemical reaction whose input nodes are called educts and whose output nodes products. In a chemical reaction, there may exist catalyzers, that is, substances that increase the rate of a reaction without being modified by it. Formally, they should be counted as both input and output nodes. That is, the two subsets of the nodes of a directed hypergraph, its input and output nodes, need not be disjoint. That will not constitute a problem for the formal concepts to be developed (see, for instance, [17] for a discussion of directed hyperloops and their curvature).

A directed hypergraph $H = (V, E)$ consists of a set V of nodes or vertices and a set E of ordered pairs of subsets of V, not both of them being empty, called hyperedges. For a hyperedge $e = (e_1, e_2) \in E$, $e_1 \subset V$ is the head of e, and $e_2 \subset V$ is its tail. We let $|f|$ be the number of vertices in $f \subset V$. We let $\deg_{in}(e)$ of a hyperedge be the number of hyperedges that have an input node of e as their head, and $\deg_{out}(e)$ the number of hyperedges that have an output node of e as their tail. Since an input edge might connect to more than one input node of e, input hyperedges are counted with the number of input nodes of e that they connect to, and analogously for output edges. As in (9), we then define the Forman-Ricci curvature of a hyperedge $e = (e_1, e_2)$ as [39]

$$F_\rightarrow(e) = |e_1| + |e_2| - \deg_{in}(e) - \deg_{out}(e). \tag{17}$$

(For a different definition of the Forman-Ricci curvature of a directed hypergraph, see [41].) Thus, here we count the number of inputs received through input nodes and the number of outputs produced at output nodes. As in [39], one can also define different types of Forman-Ricci curvature of a directed hyperedge by arranging inputs and outputs differently.

Similarly, as in (10), we can put

$$\daleth_\rightarrow(e) := \deg_{out}(e) - \deg_{in}(e). \tag{18}$$

Following [35], we can also define the Ollivier-Ricci curvature of a directed hyperedge via the Wasserstein distance between two probability measures associated to a directed hyperedge. As in (11) and (12), we need to define the corresponding measures μ_{in}, μ_{out}. In (11) and (12), the principle that was that the total measure 1 is evenly split among the inputs or the outputs, resp., unless we had a source or a sink. Now, there are more demands for splitting. A directed hyperedge may in general have more than one tail or head node, and at each of them, several incoming resp. outgoing hyperedges might be found, and each them may again have more than one tail or head. The principle then is to split the available measure at each step evenly among all the possible recipients. We shall explain the resulting splitting procedure for μ_{in} as the one for μ_{out} is analogous. Let the tail e_1 of the hyperedge $e = (e_1, e_2)$ have η_{in} elements. A source, that is, an element of e_1 without incoming hyperedges, gets the weight $\mu_{in}(v) = \frac{1}{\eta_{in}}$. To handle the others, we define the set $\mathcal{M}$ of masses of $e = (e_1, e_2)$ as the union of the tails of hyperedges that come in at an element $w \in e_1$, that is, have w in their head set. We then first divide the measure $\frac{1}{\eta_{in}}$

10

that we can distribute at such a w evenly among all those incoming hyperedges, and for each such hyperedge, we divide the measure associated with it in that manner evenly in its tail. In that manner, we assign a measure to every element in $\mathcal{M}$. Thus, we have distributed the total measure 1 among the sources and the masses of our hyperedge. This yields μ_{in}, and as mentioned, μ_{out} is constructed analogously by assigning measures to the sinks, that is, those members of e_2 and the holes, that is, the heads of hyperedges that have an element of e_2 in their tail set. The (directed) distance $d(u,v)$ between a mass u and a hole v of a hyperedge $e = (e_1, e_2)$ is defined as the minimal number of directed hyperedges connecting them. Again, it is at most 3, and this value is attained if $u \to e_1$, $e_2 \to v$ and there is no shorter way to move from u to v than to go through e. It is 0 when $u = v$ is at the same time a mass and a hole of e, and it is 1 if u is an input of a hyperedge and v is an output. Again, formally, we want to solve an optimal transport problem for moving the first probability measure to the second one. We thus minimize

$$\sum_{u \to e_i} \sum_{e_j \to v} d(u,v) \mathcal{E}(u,v) \tag{19}$$

over the set of all matrices $\mathcal{E}$ (transport plans) whose entries $\mathcal{E}(u,v)$ represent the amount of mass from $\mu_{\mathcal{M}}(u)$ moved from vertex u to vertex v.

If m_δ is the amount of mass that is moved at distance δ in an optimal transport plan, the directed Ollivier-Ricci curvature of e is defined as in (12) and becomes again as in (13)

$$O_\to(e) = m_0 - m_2 - 2m_3. \tag{20}$$

It is bounded above by 1. This is reached when $m_0 = 1$, i.e., when each mass coincides with a hole of the same size. It is bounded below by -2, reached when $m_3 = 1$, i.e., when there are no shortcuts available and each mass has to be moved through e to reach a hole. Again, (20) can be taken as the definition of $O_\to(e)$. While it depends on identifying an optimal transport plan, the formula as such is obviously very simple. For applications, see [42].

For instance, we can consider the red hyperedge in Fig. 6. Bullets represent vertices. The green bullet in the left is a source since it has no incoming hyperedge while the blue bullet in the right is a sink since it has not outgoing hyperedges. For representing masses and holes we use triangles and squares respectively. As the red hyperedge has two vertices in its tail set and each of them has at most one vertex as an incoming neighbour, the size of the masses is $1/2$. In contrast, the sizes of the four holes are different. The biggest one is the sink, with mass $1/2$. Another hole with the size $1/4$ is the top middle vertex which already got $1/2$ of the total mass. The size of the remaining hole is $1/4$, divided equally among the two vertices in the top right of the figure. Thus, both the triangles and the squares have total size 1, and the task now is to move the triangles to the squares with least total cost. There are two optimal plans, leading to the negative curvature value -1/4; In one plan, $m_0 = 1/4$, $m_1 = 1/2$, $m_2 = 0$ and $m_3 = 1/4$, while in the other $m_0 = 1/4$, $m_1 = 1/4$, $m_2 = 1/2$ and $m_3 = 0$. Note

11

that there are also transfer plans with $m_2 = 1$, and all other values 0, but they are not optimal.

These two different curvature notions represent complementary tools for detecting local geometry and connectivity patterns in directed hypergraphs. Forman curvature monotonically decreases with the number of incoming and outgoing neighbours of input and output nodes, resp., and it therefore detects hyperedges joining highly connected nodes. Ollivier curvature, on the other hand, is controlled by the overlap of the set of masses and holes (e.g. directed triangles) and by shortcuts between them (e.g. directed quadrangles and pentagons). We illustrate these principles in Fig. 4. We want to evaluate the curvatures of the black hyperedge in the left and the right figure in various constellations. Without any of the colored hyperedges, $F_\rightarrow(e) = |e_i| + |e_j|$, while $O_\rightarrow(e) = 0$. When the red edges are present, we get $F_\rightarrow(e) = 0$ in both the left and the right figure, whereas $O_\rightarrow(e)$ is negative in the left case, because there are no shortcuts, but positive in the right case, when the inputs of the tail coincide with the outputs of the head. The presence of the blue edges on the right, however, makes a difference for $F_\rightarrow(e)$, but not for $O_\rightarrow(e)$, that is, the former, but not the latter distinguishes between those cases. In contrast, while $F_\rightarrow(e)$ does not distinguish between the presence of the blue and the green edges in the right figure, $O_\rightarrow(e)$ sees the effect, as it is more negative in the presence of the green than in that of blue edges (the blue edges contribute to m_1, but the green ones to m_2).

Metabolic networks

Metabolic networks are evocative examples of directed hypergraphs, where metabolites react with others to produce products. Both, reactants, e_1, and products, e_2, typically contain more than one substance ($|e_1| \geq 1$ or $|e_2| \geq 1$) and the reactions may not be reversible. This directed relationship between sets is therefore naturally modelled by a directed hyperedge ($e_1 \rightarrow e_2$). Since metabolic networks have been extensively studied, they present an ideal setting to illustrate how to use the hypergraph tools described here. For this, let us consider the metabolic network of *Mycobacterium tuberculosis* H37Rv (version iNJ661) [43] modelled as a directed hypergraph. This network contains 939 reactions and 743 metabolites, of which 256 are reversible. Each reversible reaction ($e_1 \leftrightarrows e_2$) was divided into two, a forward reaction ($e_1 \rightarrow e_2$), and its reverse reaction ($e_1 \leftarrow e_2$). As a result, the network contains 1195 directed hyperedges. Most substrates of this metabolic network are consumed or produced by one reaction only. Also, a few are involved in more than half of the reactions ($\sim 50\%$ require h, h2o, atp or nadhp, and $\sim 57\%$ produce h, pi, h2o, adp, or co2), the distributions of indegree and outdegree in Figure 6 a) summarize this behavior. Suppose that we want to investigate whether starting materials that are produced in several different ways (large indegree in e_1) produce substances that also serve as starting materials for many reactions (large outdegree in e_2), that is, whether targets are transformed into key precursors. There are two aspects relevant for this question. First, we must find out if the network is assortative.

12

Since it is a hypergraph, we use the degree difference and its distribution shows that this is mostly the case (see Figure 5 b)). Notoriously, the degree difference is ~ 0 for 217 ($\sim 18\%$) reactions. Second, we have to locate which of those 217 reactions involve metabolites of large degree. For that, instead of looking at the difference between out- and indegrees, we need to look at their sum and turn to the distribution of $F_\rightarrow(e)$. Figure 5 a) shows that the dominant mode is represented by curvature around zero. There are also secondary humps associated with more negative curvature values. Perhaps the most important reactions, however, are those that have very low (negative) curvature values, but a degree difference near zero. In fact, the first reaction on the list is the fundamental reaction that creates the energy storage molecule adenosine triphosphate (ATP), e: `adp+h+pi` $\rightarrow$ `atp+h+h2o`, with $F_\rightarrow(e) = -1347$ and $\daleth_\rightarrow(e) = 1$. Furthermore, the associated mass set $\mathcal{M}$ shows that there are 400 precursors for the substrates of this reaction, and, based on the set of holes $\mathcal{H}$, there are 464 derived metabolites. This pair of values correspond to the upper right blue mark in Figure 6 c). Notice that this information is not given by node degree. With few exceptions, precursors and derivatives are at distances shorter than three, and mostly around zero, as shown by $O_\rightarrow(e)$ (see Figure 5 c)). For the reaction discussed here, $O_\rightarrow(e) = 0.35$ and it corresponds to the the left most blue mark of Figure 6 d). The preceding already illustrates how a combination of the three measures that we have developed, $F_\rightarrow(e), O_\rightarrow(e)$ and $\daleth_\rightarrow(e)$, can reveal the fundamental structural properties of specific reactions inside the metabolic network. Both evaluating the statistical distributions of these three quantities and comparing them for different networks, and analyzing those reactions that produce particularly prominent values for them in more detail should yield deeper insight into the structure of metabolic networks.

Acknowledgment

References

[1] Ferdinand De Saussure. Cours de linguistique générale (1916). *Edition critique préparée par T. de mauro. paris: payothèque*, 1972.

[2] Claude Lévi-Strauss. *Anthropologie structurale*, volume 171. Plon Paris, 1958.

[3] Mark EJ Newman. Mixing patterns in networks. *Physical review E*, 67(2):026126, 2003.

[4] Matthew Jackson. *Social and economic networks.* Princeton Univ.Press, 2008.

[5] Mahendra Piraveenan, Mikhail Prokopenko, and Albert Y Zomaya. Classifying complex networks using unbiased local assortativity. In *ALIFE*, pages 329–336, 2010.

[6] Amirhossein Farzam, Areejit Samal, and Jürgen Jost. Degree difference: A simple measure to characterize structural heterogeneity in complex networks. *arXiv preprint arXiv:2008.10751*, 2020.

[7] Mark Newman. *Networks.* Oxford Univ. Press, 2010.

[8] Ernesto Estrada. *The structure of complex networks.* Oxford Univ. Press, 2012.

[9] R. P. Sreejith, K. Mohanraj, J. Jost, E. Saucan, and A. Samal. Forman curvature for complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 6:063206, 2016.

[10] 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.

[11] R. P. Sreejith, J. Jost, E. Saucan, and A. Samal. Systematic evaluation of a new combinatorial curvature for complex networks. *Chaos Solitons and Fractals*, 101:50–67, 2017.

[12] Emil Saucan, Areejit Samal, Melanie Weber, and Jürgen Jost. Discrete curvatures and network analysis. *MATCH Communications in Mathematical and in Computer Chemistry*, 80(3):605–622, 2018.

[13] Areejit Samal, R. P. Sreejith, Jiao Gu, Shiping Liu, Emil Saucan, and Jürgen Jost. Comparative analysis of two discretizations of Ricci curvature for complex networks. *Scientific Reports*, 8:8650, June 2018.

[14] Emil Saucan, R. P. Sreejith, R. P. Vivek-Ananth, Jürgen Jost, and Areejit Samal. Discrete Ricci curvatures for directed networks. *Chaos, solitons and fractals*, 118:347–360, 2019.

[15] Deryc T. Painter, Bryan C. Daniels, and Jürgen Jost. Network analysis for the digital humanities : principles, problems, extensions. *Isis : a journal of the History of Science Society*, 110(3):538–554, 2019.

[16] Emil Saucan, Areejit Samal, and Jürgen Jost. A simple differential geometry for networks and its generalizations. In Hocine Cherifi, Sabrina Gaito, José Fernendo Mendes, Esteban Moro, and Luis Mateus Rocha, editors, *Complex networks and their applications VIII : Volume 1 Proceedings of the eighth international conference on complex networks and their applications ; Complex networks 2019*, volume 881 of *Studies in computational intelligence*, pages 943–954. Springer, Cham, 2020.

[17] Wilmer Leal, Marzieh Eidi, and Jürgen Jost. Ricci curvature of random and empirical directed hypernetworks. *Applied network science*, 5:65, 2020.

[18] C. Berge. *Graphs and Hypergraphs*. Elsevier Science Ltd., Oxford, UK, UK, 1985.

[19] Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2):177 – 201, 1993.

[20] Gourab Ghoshal, Vinko Zlatić, Guido Caldarelli, and M. E. J. Newman. Random hypergraphs and their applications. *Phys. Rev. E*, 79:066118, 2009.

[21] Alain Bretto. *Hypergraph Theory: An Introduction*. Springer Publishing Company, Incorporated, 2013.

[22] Cliff A. Joslyn, Sinan Aksoy, Tiffany J. Callahan, Lawrence E. Hunter, Brett Jefferson, Brenda Praggastis, Emilie A. H. Purvine, and Ignacio J. Tripodi. Hypernetwork science: From multidimensional networks to computational topology. *arXiv preprint arXiv:2003.11782*, 2020.

[23] Carl Adam Petri. Fundamentals of a theory of asynchronous information flow. In *IFIP congress*, volume 62, pages 386–390, 1962.

[24] Carl Adam Petri and Wolfgang Reisig. Petri net. *Scholarpedia*, 3(4):6477, 2008.

[25] Raffaella Mulas, Christian Kuehn, and Jürgen Jost. Coupled dynamics on hypergraphs : master stability of steady states and synchronization. *Physical review / E*, 101(6):062313, 2020.

[26] Anirban Banerjee and Samiron Parui. On synchronization in dynamical hyper-networks. *arXiv preprint arXiv:2008.00469*, 2020.

[27] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri. Networks beyond pairwise interactions: structure and dynamics. *Physics Reports*, 2020.

[28] Guilherme Ferraz de Arruda, Giovanni Petri, and Yamir Moreno. Social contagion models on hypergraphs. *Physical Review Research*, 2(2):023032, 2020.

[29] Telmo Menezes and Camille Roth. Semantic hypergraphs. *arXiv preprint arXiv:1908.10784*, 2019.

[30] Jürgen Jost. *Riemannian geometry and geometric analysis*. Universitext. Springer, Berlin [u. a.], seventh edition edition, 2017.

[31] Frank Bauer, Bobo Hua, Jürgen Jost, Shiping Liu, and Guofang Wang. The geometric meaning of curvature: Local and nonlocal aspects of ricci curvature. In Laurent Najman and Pascal Romon, editors, *Modern Approaches to Discrete Curvature*, pages 1–62. Springer International Publishing, Cham, 2017.

[32] Robin Forman. Bochner's method for cell complexes and combinatorial Ricci curvature. *Discrete and Computational Geometry*, 29:323–374, 2003.

[33] David N Fisher, Matthew J Silk, and Daniel W Franks. The perceived assortativity of social networks: methodological problems and solutions. In *Trends in Social Network Analysis*, pages 1–19. Springer, 2017.

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

[35] Marzieh Eidi and Jürgen Jost. Ollivier ricci curvature of directed hypergraphs. *Scientific Reports*, 10(1):12466, 2020.

[36] Jürgen Jost and Shiping Liu. Ollivier's ricci curvature, local clustering and curvature-dimension inequalities on graphs. *Discrete and Computational Geometry*, 51(2):300–322, 2014.

[37] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[38] Tommy V Vo, Jishnu Das, Michael J Meyer, Nicolas A Cordero, Nurten Akturk, Xiaomu Wei, Benjamin J Fair, Andrew G Degatano, Robert Fragoza, Lisa G Liu, et al. A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. *Cell*, 164(1-2):310–323, 2016.

[39] Wilmer Leal, Guillermo Restrepo, Peter F. Stadler, and Jürgen Jost. Forman-Ricci curvature for hypergraphs. *arXiv preprint arXiv:1811.07825*, 2018.

[40] Kyle J Minch, Tige R Rustad, Eliza JR Peterson, Jessica Winkler, David J Reiss, Shuyi Ma, Mark Hickey, William Brabant, Bob Morrison, Serdar Turkarslan, et al. The dna-binding network of mycobacterium tuberculosis. *Nature communications*, 6:5829, 2015.

[41] Emil Saucan and Melanie Weber. Forman's Ricci curvature – from networks to hypernetworks. Technical Report 1810.07749v1, ArXiv, 2018.

[42] Wilmer Leal, Marzieh Eidi, and Jürgen Jost. Curvature-based analysis of directed hypernetworks. In Hocine Cherifi, editor, *Complex networks 2019 : the 8th international conference on complex networks and their applications ; December 10 - 12, 2019 Lisbon, Portugal ; book of abstract*, pages

32–34. International Conference on Complex Networks and Their Applications, [s.l.], 2019.

[43] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007.
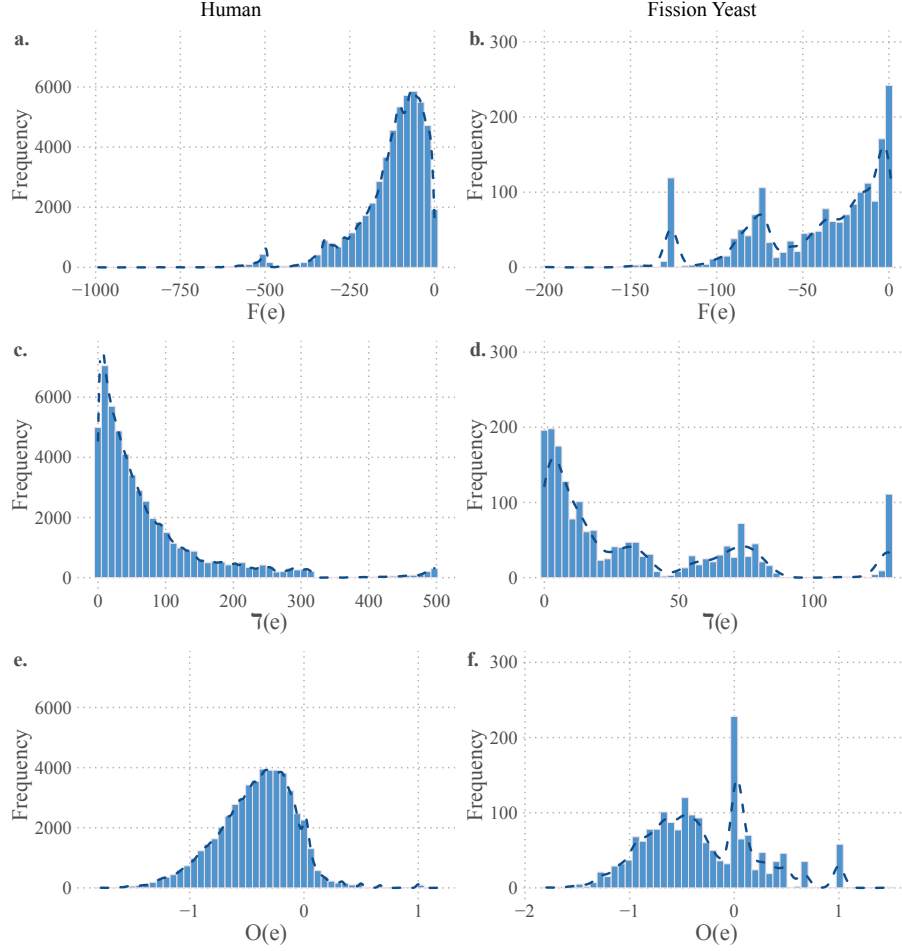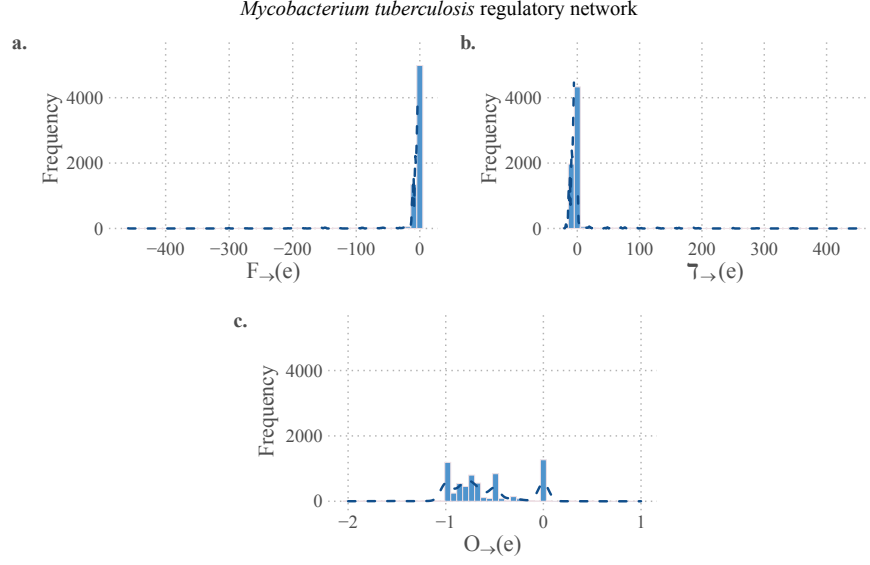
Figure 1: The distribution of **(a,b)** Forman-Ricci curvature, **(c,d)** degree difference, and **(e,f)** Ollivier-Ricci curvature in the giant components of the binary protein interaction networks in human (left) and fission yeast (right), respectively. In each case, protein-protein interactions are represented via an undirected and unweighted graph. The nodes and edges represent proteins and binary interactions between them, respectively. The giant component of the human network has 8152 nodes and 52036 edges, while of the fission yeast network has 1306 nodes and 2278 edges.

Figure 2: The distribution of **(a)** Forman-Ricci curvature, **(b)** degree difference, and **(c)** Ollivier-Ricci curvature in the transcriptional regulatory network of *Mycobacterium tuberculosis*. There are 6581 unweighted directed edges and 2547 unweighted nodes. The source in each directed edge is a transcription factor (TF) and the target is a target gene controled by the TF.
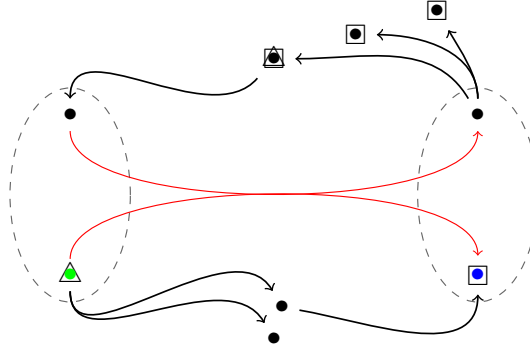

Figure 3: The red hyperedge is negatively curved as in an optimal transference plan, the size of coincident masses (triangles) and holes (quadrangles), located on the top middle vertex, is less than the size of the masses which need to be moved with distance 2. Also the two colored vertices in the left and the right of the figure are a source and a sink since they have no incoming resp. outgoing hyperedges.
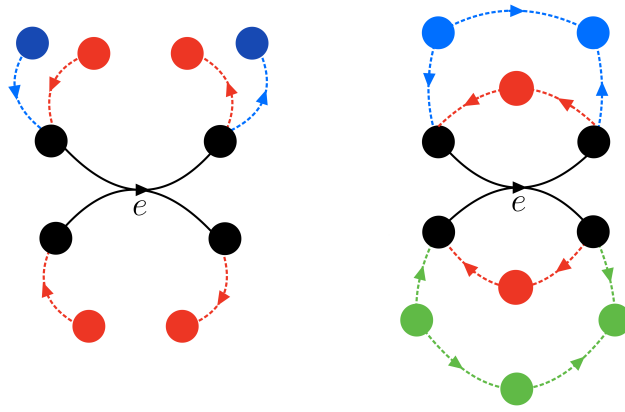
Figure 4: Illustration of the different connectivity patterns that affect $F(e)$ and $O(e)$. Well-connected hyperedges often play a key role in a network. Said hyperedges are identified by $F_{\rightarrow}(e) = |e_1| + |e_2| - \deg_{in}(e) - \deg_{out}(e)$, since it decreases monotonically with the number of incoming neighbors to the tail and outgoing neighbors from the head of e. Nevertheless, $F_{\rightarrow}(e)$ is not affected by the presence of arcs from the former to later. $O_{\rightarrow}(e) = m_0 - m_2 - 2m_3$ captures this complementary information. On the right, we find directed triangles, which contribute to m_0 (black and red hyperedges), directed quadrangles to m_1 (black and blue hyperedges), and directed pentagons to m_2 (black and green hyperedges). In the figure on the left, the shortest path between any incoming and any outgoing neighbor, is 3. Such a connectivity pattern contributes to m_3.
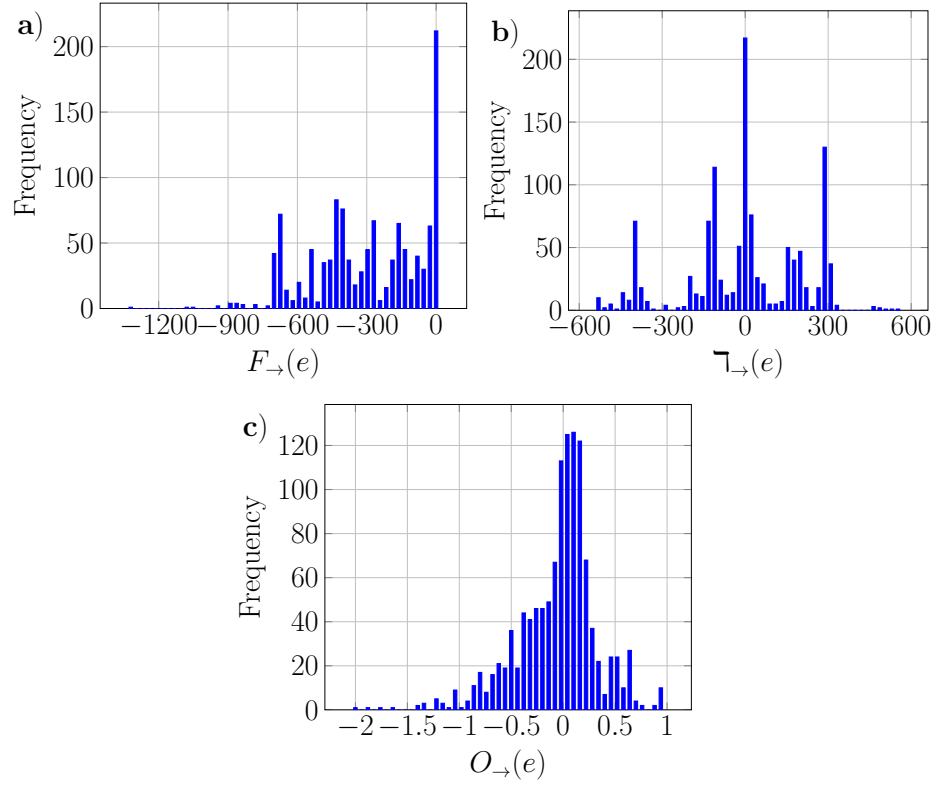
Figure 5: The distribution of (a) Forman-Ricci curvature, (b) degree difference, and (c) Ollivier-Ricci curvature in the metabolic network of *Mycobacterium tuberculosis* H37Rv, which is represented as a directed hypergraph with nodes as *M. tuberculosis* metabolites and directed hyperedges as chemical reactions. The network has 743 nodes and 1195 hyperedge edges.
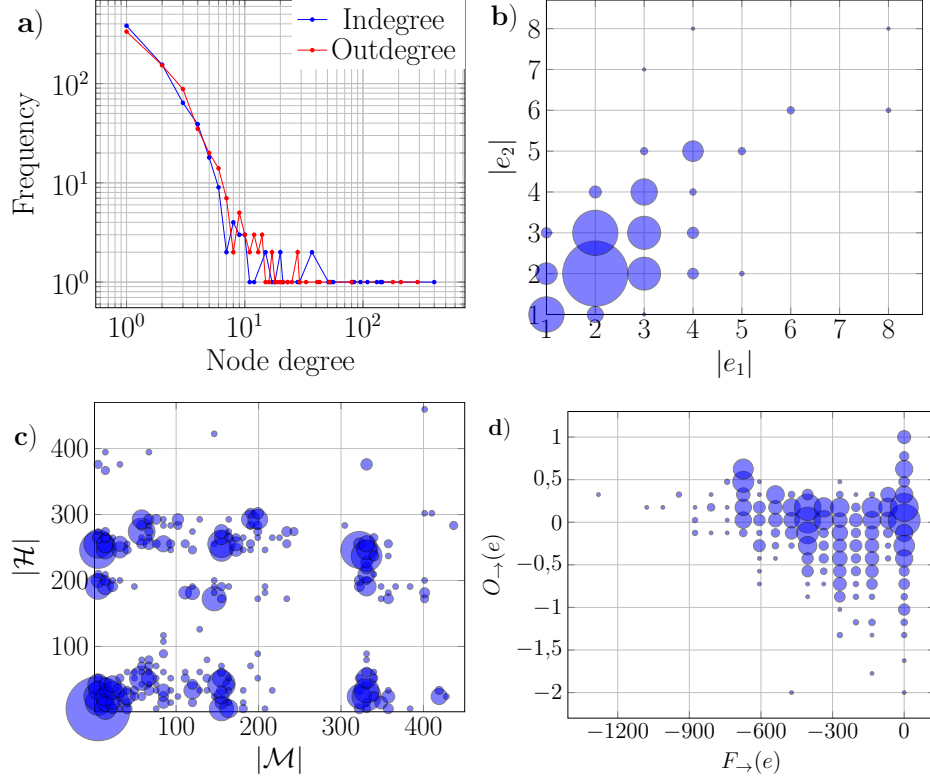
Figure 6: The distribution of (a) node degree, (b) sizes of tails and heads, $(|e_1|, |e_2|)$, (c) number of masses and holes, $(\mathcal{M}, \mathcal{H})$, and (d) $(\mathrm{F}(e), \mathrm{O}(e))$ values in the metabolic network of *Mycobacterium tuberculosis* H37Rv, which is represented as a directed hypergraph with nodes as *M. tuberculosis* metabolites and directed hyperedges as chemical reactions. The network has 743 nodes and 1195 hyperedge edges.