Hypergraph topological quantities for tagged social networks

Vinko Zlatić,[1,2] Gourab Ghoshal,[3] and Guido Caldarelli[1,4]

[1]*CNR-INFM Centro SMC Dipartimento di Fisica, Università di Roma "Sapienza," Piazzale A. Moro 5, 00185 Roma, Italy*
[2]*Theoretical Physics Division, Rudjer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia*
[3]*Department of Physics and Michigan Center for Theoretical Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*
[4]*Linkalab, Complex Systems Computational Laboratory, 09100 Cagliari, Italy*

Recent years have witnessed the emergence of a new class of social networks, which require us to move beyond previously employed representations of complex graph structures. A notable example is that of the folksonomy, an online process where users collaboratively employ tags to resources to impart structure to an otherwise undifferentiated database. In a recent paper, we proposed a mathematical model that represents these structures as tripartite hypergraphs and defined basic topological quantities of interest. In this paper, we extend our model by defining additional quantities such as edge distributions, vertex similarity and correlations as well as clustering. We then empirically measure these quantities on two real life folksonomies, the popular online photo sharing site Flickr and the bookmarking site CiteULike. We find that these systems share similar qualitative features with the majority of complex networks that have been previously studied. We propose that the quantities and methodology described here can be used as a standard tool in measuring the structure of tagged networks.

I. INTRODUCTION

Over the past decade, the recognition of complex networks as a useful and versatile mathematical representation of various real world systems, has led to a huge volume of work, studying its topological and dynamical properties [1–6]. A variety of models have been proposed, ranging from those describing simple undirected graphs—a basic representation of a communication network for example—to more complicated bipartite networks representing collaboration networks such as board of directors in a company, or movie actors— see [7–9].

However, the advent of Web 2.0 and its associated new forms of user-driven content have led to new social systems that cannot be adequately described by existing models. One such example is related to a phenomenon known as *folksonomy* [10,11]. In this process, users collaboratively create and manage *tags* to categorize and annotate data. Unlike traditional forms of data indexing, where administrators of a particular webpage maintain and categorize the content, in a folksonomy, both creators and consumers are free to participate in the process. Instead of a controlled set of keywords, tagging networks consist of a user generated taxonomy.

Consider the example of the popular file-sharing database known as Flickr. In this website, users can create an account and upload their personal photos. In addition to uploading photos, they are free to give them a short text description using tags. These photos (in most cases) can then be viewed by other users, who in turn can assign additional tags to the photo depending on their preferences, and so the process continues. There are also a number of other websites of a similar nature, but dealing with different resources. For instance, in the website CiteUlike, users upload and assign tags to academic papers as opposed to photographs.

Roughly speaking, tagged networks can be divided into two categories. In the first case, users are presented with a variety of available key words, which they can then freely employ to resources of their choice. Although this represents a degree of control in the set of tags that are available to users, the mechanism by which this control arises is still decentralized. In Flickr, when a user uploads a photograph and gives it a short text description, that description or tag is always public, which is to say that anyone visiting the site can see the full set of tags describing the photograph. This serves a number of functions. On the one hand it prevents the practice of redundant tagging, since once a particular tag has been applied to a resource, one is not allowed to retag the item with the same description; on the other hand it also provides new users with a previously employed set of popular tags, which they can then use on their own photographs. Finally, if none of the previously employed tags are appropriate to newly uploaded resources, then users are forced to supply sufficiently different descriptions. In this way, the set of keywords present in the network represent a reasonably well organized and diverse set. In other websites such as Citeulike, tags are not always public, and this process of decentralized control is not present. Consequently, this may give rise to vastly different statistical properties.

Some attempts have been made recently, to try and describe these tagging systems. Among them, people have tried to model them as simple unipartite and bipartite graphs, as well as simplified forms of tripartite graphs [12–14]. In addition there have been a number of studies focusing particularly on the tags such as the definition of communities [15], clustering [16], and global measures such as PageRank [17].

The key thing to note however, is that unlike in a simple network merely consisting of vertices, and edges describing the association between them, in a tagging network the fundamental building block is a triple consisting of a *user*, a *resource* that the user uploads and finally a *tag* that the user employs to describe the resource. A complete representation of such folksonomy data must capture this three-way relationship, and this leads us to consider hypergraphs.

A hypergraph is a generalization of a regular graph in the sense that an edge can connect multiple vertices. So unlike in a regular graph, where an edge connects two vertices, in a hypergraph a hyperedge is a collection of arbitrary number of vertices. These vertices can be of the same or different types, and hyperedges can vary in the number of vertices they connect. This fits in quite nicely with how tagging networks are organized. By representing a triple, as a hyperedge, one can conveniently preserve the structure of the network and examine its properties in its entirety.

In a previous paper [18], we defined a mathematical null model that represents these folksonomies as random tripartite hypergraphs and defined some basic topological quantities of interest such as the degree distribution, and component structure. In addition we calculated a number of properties of the model in the limit of large system size. In this paper, based on the hypergraph representation, we define a number of other useful topological features, such as the edge distribution, hyperedge distribution, vertex similarity, distance measures and the clustering coefficient as well as a simple definition of community structure based on the similarity between vertices. We then measure these quantities on data sets gathered from two real folksonomies, Flickr [19], and CiteULike [20]. We find that these networks share a number of qualitative features with previously studied social networks.

II. TRIPARTITE HYPERGRAPHS

We begin our analysis of the folksonomies, by first defining the representation that we will be using. We represent the network as tripartite graphs consisting of three different types of vertices, which we will refer to as red, green, and blue. For the purposes of our study, red will represent users, blue tags, and green resources, however, the colors themselves are secular as to what they represent. The edges represent three-way hyperedges that each connect exactly one red, one green, and one blue vertex. In addition, we can also color the regular edges, depending on the types of vertices they connect. For example the edge connecting a blue and green vertex is cyan (since blue and green combine to form cyan in the visual spectrum). Similarly the other edges are colored yellow (red-green) and magenta (red-blue). This classification of different regular edges allows us to measure quantities such as, the number of hyperedges a given regular edge participates in and so on. A visual illustration of this is shown in Fig. 1.

To couch this in the language of graph theory, our representation corresponds to the case of a tripartite hypergraph $G=(V,H)$, which can be defined as a pair of sets V and H, that satisfy the following conditions: (i) the set $V =\{V_r,V_g,V_b | V_i \cap V_j = \oslash\}$ is formed by the union of three disjoint sets, and (ii) the set $H \subset \{(v_r \in V_r, v_g \in V_g, v_b \in V_b)\}$ of hyperedges is a triangle connecting elements of these three sets.

In [18], we investigated number of basic properties of such hypergraphs, such as the tripartite analog of the vertex degree, component structures and projections of the network into the space of bipartite and unipartite graphs. In addition
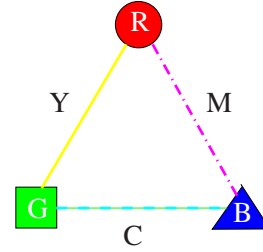

FIG. 1. (Color online) A hyperedge representing the fundamental building block in our network. Each hyperedge consists of three types of vertices, red circles, green squares, and blue triangles. In addition the regular edges are also colored according to the types of vertices they connect. In relation to folksonomies, the circles represent the users, the squares the resources, and the triangles the tags.

we defined a random graph model, related to a version of the configuration model for regular graphs [21,22] and calculated these properties exactly in the limit of large graph size. One of the assumptions made in the model was that the hypergraphs were locally treelike, in the sense that there were a trivial number of short-range loops connecting vertices. In real folksonomies however, this assumption is not strictly true, and to reflect this we extend our model by defining a number of other properties of interest to examine the loop structure as well as correlations in the network.

In particular, we measure the following quantities:

(i) *edge degrees*: defined as the number of hyperedges that a regular edge participates in. For example, a magenta edge connecting red and blue vertices might participate in a triple with a number of other green vertices. In the language of folksonomies, this represents the number of resources that a user has described with the same tag.

(ii) *clustering*: defined as the degree of overlap between the different hyperedges that a vertex participates in.

(iii) *vertex-vertex distance*: defined as the shortest paths between two nodes that are reachable along hyperedges, as well as via colored regular edges.

(iv) *community structure*: defined on the basis of vertex similarity between nodes of the same type.

A. Degrees

There are a number of options available to us when defining the degree of a vertex or edge in a tripartite graph. The simplest and most reasonable choice for a vertex is to count its degree as the number of hyperedges it participates in. Thus, a red vertex that connects to four hyperedges has degree four. The same applies to vertices of different colors. If there are H hyperedges in the network, and N_r red, N_b blue, and N_g green vertices, then the mean degree of each vertex is fixed by the condition,

$$N_r c_r = N_b c_b = N_g c_g, \qquad (1)$$

where c_r represents the mean degree of red vertices, with c_b and c_g the corresponding quantities for blue and green. (This follows from the fact that each hyperedge consists of a single red, green, and blue vertex.)

Just as in the case of regular graphs, we can define a degree distribution for each of the colors. We define $P(k_r)$ to
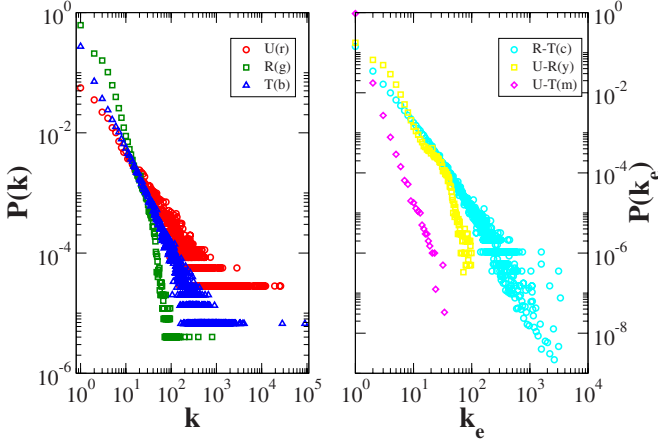
FIG. 2. (Color online) The two types of hyperdegree distributions found in a subset of the website CiteULike. Left panel: The vertex degree distributions for users, tags and resources. Right panel: The degree distribution for the various edge types.
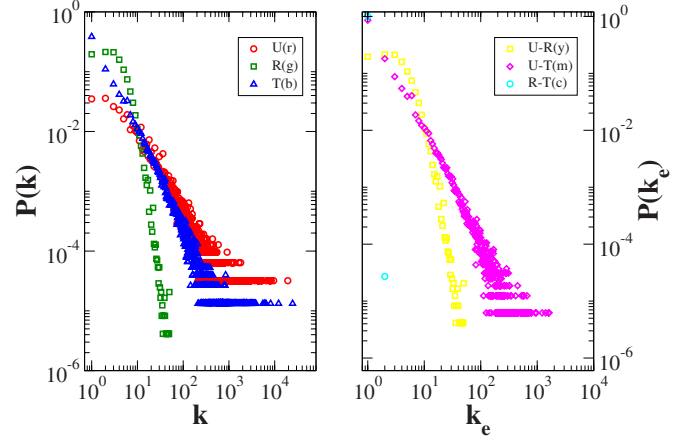

FIG. 3. (Color online) The two types of hyperdegree distributions found in a subset of the website Flickr. Left panel: The vertex degree distributions for users, tags and resources. Right panel: The degree distribution for the various edge types. The near absence of data points for the cyan edge (resource-tag) is related to the fact that tags in Flickr are public and this prevents redundant tagging—that is the application of the same tag to the same resource by multiple users.

be the fraction of red nodes in the network with hyperdegree k_r, as well as $P(k_b)$ and $P(k_g)$ corresponding to blue and green, respectively. These probability distributions satisfy the usual sum rules,

$$\sum_{k_r=0}^{\infty} P(k_r) = \sum_{k_g=0}^{\infty} P(k_g) = \sum_{k_b=0}^{\infty} P(k_b) = 1, \quad (2)$$

with $c_r = \Sigma_{k_r} k_r P(k_r)$ and similarly for the other two colors.

In addition to the degrees of vertices, we can also define corresponding quantities for regular edges. Say there are H_y number of yellow edges, then we define the degree of one of these edges as k_y—the number of different hyperedges it contributes to. We can think of these edges as representing pairs of vertices, such that in the context of folksonomies, the degree k_y corresponds to the number of different tags that a given user applies to a particular resource. The quantities k_m and k_c represent the same for the other two types of edges. In exactly the same way as for the vertices, we can define edge degree distributions thus,

$$\sum_{k_c=0}^{\infty} P(k_c) = \sum_{k_y=0}^{\infty} P(k_y), = \sum_{k=0}^{\infty} P(k_m) = 1. \quad (3)$$

with the mean degree c of each edge fixed by the condition,

$$H_c c_c = H_y c_y = H_m c_m, \quad (4)$$

where $H = H_c + H_y + H_m$.

We measure these two different quantities on data sets on our two example folksnomies, CiteUlike, and Flickr. On the left side of Figs. 2 and 3 we show the vertex degree distribution for both websites. As is fairly common for most social networks both of these show a fat-tailed distribution. On the right hand side of each figure we show the edge degree distribution. Once again these distributions are right skewed. A notable difference is the distribution of common users for a given resource-tag pair in Flickr (the cyan edge). As discussed in the introduction, this is related to the different tagging schemes in the two networks. Note that the phenomena

of multiple users applying the same tag to the same resource is representative of redundant tagging or some sort of spam. Since in CiteUlike, the tags applied by a user to a resource is not always visible to other users of the website, the incidence of users applying the same description to a paper is much higher. In Flickr, however the tags are public, and once a tag is applied to a photograph no one else is allowed to employ the same tag to that photo—thus, the near absence of any data points in the distribution of cyan edges.

Apart from the individual vertices and edges, we can also consider a hyperedge or triple as a basic unit and measure the number of other hyperedges, say h_h, that are connected to it via any of its individual constituents. In the spirit of regular unipartite graphs, one can loosely think of this as a measure of degree-degree correlations.

For a given hyperedge, this quantity can be easily computed from the individual degree of each of its constituent vertices and edges in the following manner,

$$h_h = k_r + k_g + k_b - k_c - k_m - k_y, \quad (5)$$

where the indices represent the color of the different vertices and edges. So say for example, our network just consists of two hyperedges that share a common red vertex and we look at one of the triangles and examine the number of other hyperedges it connects to. Each of the blue and green vertices have degree one, all the edges have degree one, whereas the red vertex has degree two (since it is part of two hyperedges), then Eq. (5), correctly tells us that our hyperedge has degree one.

In the same way as the vertex and edge degrees we can also define and calculate the distribution for the number of hyperedge neighbors of a given hyperedge. Let $P(h_h)$ represent the fraction of hyperedges in the network that are connected to exactly h_h other hyperedges in the sense described above. Assuming that there are no correlations between the
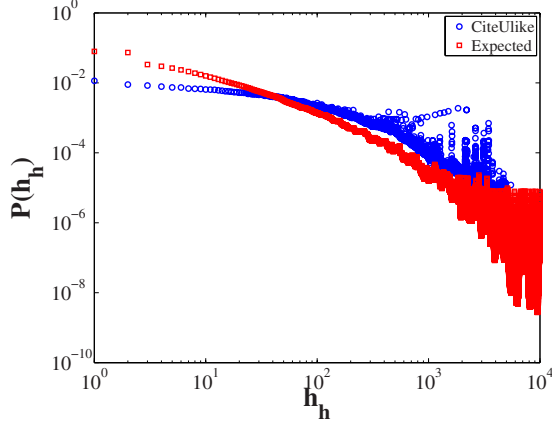
FIG. 4. (Color online) The distribution of hyperedge neighbors $P(h_h)$ as measured from the data set in CiteULike (the spikes in the data are related to spamming), as well as the theoretical values computed from Eq. (6).

degrees of the vertices and the edges, one can then define the expected probability distribution of hyperedge neighbors, using the relation

$$P(h_h) = \sum_{\substack{k_r, k_g, k_b, \\ k_c, k_m, k_y}} P(k_r) P(k_g) P(k_b) P(k_c) P(k_m) P(k_y) \cdot \Theta(k_r - k_m)$$
$$- k_y) \Theta(k_g - k_c - k_y) \cdot \Theta(k_b - k_m$$
$$- k_c) \delta_{h_h, k_r + k_g + k_b - k_c - k_m - k_y}, \quad (6)$$

where $\Theta(x)$ represents the Heaviside step function and $\delta_{x,y}$ is the Kronecker delta. The equation above assumes that the probability distribution for each vertex/edge type are statistically independent. In Figs. 4 and 5 we show the distribution of measured hyperedge degrees from the data sets (blue circles) as well as the values predicted by Eq. (6) (red circles). As is clearly visible the agreement between the two curves for both Flickr and CiteUlike is at great variance, thus suggesting that the degree of the vertices as well as the edges
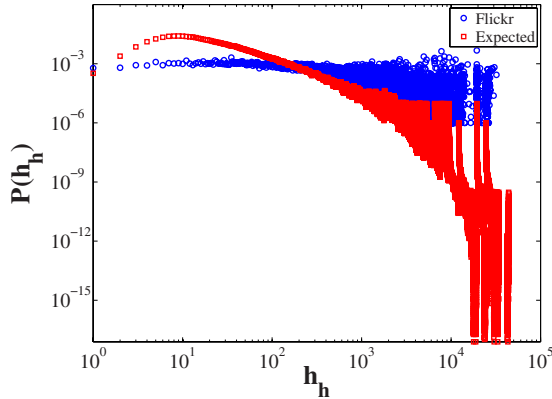

FIG. 5. (Color online) The distribution of hyperedge neighbors $P(h_h)$ as measured from the data set in Flickr as well as the theoretical values computed from Eq. (6). It is evident that the assumption of statistical independence between vertex and edge degrees is not appropriate in this case.

are correlated in some fashion and cannot be treated independently.

B. Clustering

As is well known, many networks show a high degree of clustering or transitivity—the tendency of two neighbors of a given vertex to also be neighbors of each other—thus, forming triangles of connections. The average of the probability of such types of connections is called the clustering coefficient. It is instructive to determine if this effect is predominant in folksonomies.

Once again there are a number of ways to define clustering in tripartite graphs. For example, one can project the graph onto the space of a vertex of a particular color (say that of users [18]), and then use the standard measure of clustering. However, as discussed before, our aim is to keep the tripartite structure of our network intact, and therefore we will define an analog of clustering that takes into account the full three-way relationship between the vertices.

As motivation for our definition of clustering, consider a red vertex (a user), that is connected to three hyperedges. If the graph was locally treelike, then this would imply that the red vertex connects to three-blue and three-green neighbors. However it is possible that some of its neighbors might be common to more than one hyperedge, thus, the number of blue and green neighbors could be less than three. One example is if a user assigns three tags to the same resource, then it has three tags as neighbors and only one resource. Thus, one can think of this measure of overlap between different hyperedges as a close analog to clustering for regular graphs, in the sense that it is a metric for the deviation of the network from being treelike.

In order to quantify this measure, we first define the coordination number z for a given vertex, as the number of immediate neighbors of any color that are connected to it via regular edges (this is just the standard definition of degree for unipartite graphs). For a vertex with k hyperedges one can define upper and lower bounds for the coordination number. If there were no overlap, that is to say, the vertex shares no common neighbors between its k hyperedges, then the maximal coordination number z_{max} is equal to $2k$, since it is connected to two other vertices via each hyperedge. One can show that in the case of maximum overlap, the corresponding expression for z_{min} is $z_{min}(k) = 2n$ for $n(n-1) < k \le n^2$ and $z_{min} = 2n+1$ for $n^2 < k \le n(n+1)$, with n some integer.

Based on the coordination number defined above for a vertex i of degree k_i, we define a local measure of overlap or clustering, the *hyperedge density* $D_h(i)$ of vertex i as:

$$D_h(i) = \frac{z_{max} - z(i)}{z_{max} - z_{min}}. \quad (7)$$

It can be immediately seen that if a vertex does not share any common neighbors between its hyperedges, then $z = z_{max}$, and the hyperedge density vanishes, which the means the neighborhood of the vertex is locally treelike. In the case of maximum overlap, $z = z_{min}$, and the ratio is then $D_h = 1$. One can also define an average hyperedge density over all vertices with degree k thus,
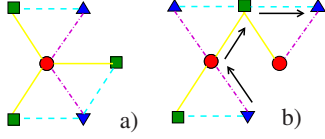
FIG. 6. (Color online) In a) the central red node has $D_h=0.5$ as per Eq. (7), because $z_{max}=6$ and $z_{min}=4$. In b) a path of distance 3 (marked by arrows) between two tags via a tag-user, user-resource, and resource-tag path.

$$D_h(k) = \frac{\Sigma_i D_h(i) \delta_{k_i,k}}{\Sigma_i \delta_{k_i,k}}. \tag{8}$$

In Fig. 6(a), we show an example analysis of the hyperedge density.

In Fig. 7, we show the measurement of the average hyperedge density of vertices with degrees k degree for both our example websites. The plot shows that both CiteULike and Flickr share a high incidence of overlap between its hyperedges. In fact the value of D_h is generally always larger than 0.5, which is to suggest connections of the type shown in Fig. 6(a) are fairly common. For both types of networks, the hyperedge density of users is significantly larger than that of resources or tags. There is possibly a fairly simple explanation for this. In Flickr for example, users typically upload photographs in sets and then apply descriptive tags to the same set. Thus, many different resources share similar tags associated with the same user. Therefore although a user might participate in a number of hyperedges a majority of them are associated with either a common resource that has been assigned multiple tags, or a common tag that has been used to describe a number of resources by the same user. The lower hyperedge density values for individual tags imply that they are employed by a large variety of different user-
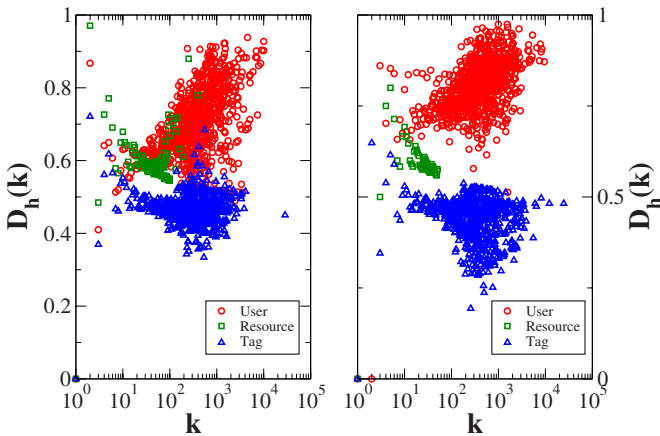
resource pairs. In the case of Flickr, this might be reflective of the more diverse and ordered set of tags that arise due to the decentralized control described in the introduction.

C. Vertex-Vertex Distance

Another important quantity of interest is the average distance between a pair of vertices in a graph. This is important for a number of reasons. One application is related to searching for resources in a network. In CiteUlike for example, a user might be interested in looking for a particular paper. In order to do so one would have to *surf* the hypergraph, through the various hyperedges. In some sense the efficacy of the search is related to how far apart vertices of different types are in the network. For example, one might find that surfing on the network of tags would lead one to a resource in much faster time, than if one were to look through the list of users. The same considerations apply to automated *web crawlers* that crawl through websites to perform directed searches, or to create indexes for later search. The knowledge of the distance between various types of vertices in the network can lead to more efficient paths being chosen and thus more effective search schemes.

In the case of tripartite-hypergraphs, the shortest paths between different vertices can be defined as the minimal number of hyperedges, which connect those vertices. This definition follows from the definition of shortest paths in ordinary graphs. The flow of information through the hyperedges can be simply described as a hopping process along vertices sharing a common hyperedge. In addition to this, it is also interesting to measure the paths through the differently colored regular edges. As mentioned before this might

FIG. 7. (Color online) The average hyperedge density $D_h(k)$ as a function of the degree k of a vertex. On the left panel, the hyperedge density for different vertex categories in the CiteULike data set. On the right panel, the same quantities for Flickr. For both data sets the users tend to form increasingly dense connections with increasing k. The density for resources seem to fall off with increasing k, except in CiteULike, where there is an interesting turning point around $k=100$.

FIG. 8. (Color online) The distribution of paths between different types of vertices, implying that the network as a whole exhibits the so called small- world effect. The black arrow shows that the number of user-tag paths of distance 2 are comparatively few. This means that there are very few resources that function as direct bridges between a user and a tag. The red arrow shows that are fewer paths between tags of length 3, as compared to those of length 2 and 4. This implies that the number of bridging hyperedges where two tags are connected via a user and then a resource (as shown in Fig. 6(b)) are rare in this network.

FIG. 9. (Color online) The network of tags in CiteUlike, constructed via the application of the similarity measure shown in Eq. (9)—in this case considering the set of papers as neighbors. Vertices with distance larger than 0.9 have been discarded. The connection between tags "differential" and "equation" indirectly implies the clustering of the corresponding papers into the same group.

help in defining an efficient hopping scheme for an automated crawler, which could try differently colored paths depending on which one is closest to a desired target at each step of the crawl. An example of the distance between two vertices (tags in this case) based on the hyperedges and the regular colored edges is shown in Fig. 6(b).

We took a subset of the data from the website CiteUlike (denoted CiteManageable), and measured the average distance between vertices of the same type as well those of different types. The results are presented in Fig. 8. In all cases, it seems that the average distance peaks around paths of length four, which suggests that the network as a whole exhibits the small-world effect [23] also present in a variety of other networks.

D. Community Structure

A question of particular importance, is to examine whether our example folksonomies exhibit community structure—the tendency of the network to divide naturally into groups of nodes with dense connections within groups and sparser connections between groups. For example there might be different groups of users in the network that share commonality with themselves owing to similar tastes in content—in Flickr users who share an interest in pictures of art—or one might find groups of tags that occur together many number of times, say *differential*, *equation*, and *series expansion*, when describing physics papers in the website CiteULike.

Since most of the data sets (of significance) available from both Flickr and CiteUlike are fairly large (on the order of a million hyperedges at the very least), we employ an approach based on a local quantity—vertex similarity. The vertex similarity is a measure of vertex "distance" defined as

$$\rho(v_1, v_2) = \frac{(N_1 \cup N_2) - (N_1 \cap N_2)}{(N_1 \cup N_2) + (N_1 \cap N_2)}, \quad (9)$$

where N_1 and N_2 are neighbors of the vertices v_1 and v_2, respectively. The numerator is the standard Euclidean—or Hamming distance in information theory—and the denominator is the just the sum of the degrees. Note that the measure can meaningfully be employed only to vertices of the same type (similar tags for example), and not necessarily to those of unlike types. In addition, for tripartite graphs, one has to chose the type of neighbor. So, if we were to look at the similarity between two tags, we can either consider its set of neighboring users or neighboring resources. This approach is particularly useful for social networks like folksonomies, since it leads to self-categorization of content, via a bottom-up procedure (which seems natural for such decentralized systems).

In Fig. 9, we show the result of the application of this to a subset of the data taken from CiteUlike. The figure shows groups of similar tags, where we used the papers as the neighbor set—in other words two tags are similar if they have employed many times together to different papers. In

this particular example, tags are connected if their distance $\rho(v_1, v_2)$ is lower than a given threshold, $(\rho(v_1, v_2) < \leq 0.9)$. In principle one can tune these connected structures by modifying the value of this threshold parameter. As the figure shows tags such as *differential* and *equation* indeed are similar to each other in the sense considered here.

III. CONCLUSIONS

In this paper, we have examined the structural properties of two types of social networks, so called folksonomies, consisting of users applying descriptive tags to resources. In order to preserve this three-way relationship, we have represented this structure as a tripartite hypergraph, with a user-resource-tag triple representing a hyperedge.

We have defined a number of topological quantities of interests, such as a variety of degree distributions, correlations, clustering, distance distributions as well as a simple metric for discerning community structure. We then empirically measured these quantities on data taken from subsets of our example networks, Flickr and CiteUlike. We find that these networks share a number of qualitative features with previously studied social networks such as the presence of fat tails in the statistical distributions of links, the small world property in terms of the distance between vertices, as well as a high degree of clustering.

We propose that the topological measures as well as the methodology proposed here can be used as a standard tool for measuring the properties of networks of a similar nature.

ACKNOWLEDGMENTS

[1] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).

[2] M. Faloutsos, P. Faloutsos, and C. Faloutsos, in SIGCOMM, Comput. Commun. Rev. **29**, 251 (1999).

[3] G. Bonanno, G. Caldarelli, F. Lillo, and R. N. Mantegna, Phys. Rev. E **68**, 046130 (2003).

[4] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).

[5] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).

[6] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[7] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[8] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks From Biological Nets to the Internet and the WWW* (Oxford University Press, Oxford, 2003).

[9] G. Caldarelli, *Scale-Free Networks* (Oxford University Press, Oxford, 2007).

[10] C. Cattuto, V. Loreto, and L. Pietronero, Proc. Natl. Acad. Sci. U.S.A. **104**, 1461 (2007).

[11] O. Görlitz, S. Sizov, and S. Staab, in Proceedings of ESWC-2008, European Semantic Web Conference, LNCS, Lect. Notes Comput. Sci. **5021**, 807 (2008).

[12] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme, AI Comm. **20**, 245 (2007).

[13] R. Lambiotte and M. Ausloos, Lect. Notes Comput. Sci. **3993**, 1114 (2006).

[14] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, and T. Vicsek, New J. Phys. **10**, 123026 (2008).

[15] L. Specia and E. Motta, Lect. Notes Comput. Sci. **4519**, 624 (2007).

[16] A. Capocci and G. Caldarelli, J. Phys. A **41**, 224016 (2008).

[17] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme, Lect. Notes Comput. Sci. **4011**, 411 (2006).

[18] G. Ghoshal, V. Zlatić, G. Caldarelli, and M. E. J. Newman, Phys. Rev. E **79**, 066118 (2009).

[19] Flickr data can be downloaded at https://www.uni- koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets.

[20] CiteULike data can be downloaded as explained in http://www.citeulike.org/faq/data.adp.

[21] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).

[22] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).

[23] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).