A hypergraph model of social tagging networks

J. Stat. Mech. (2010) P10005

(http://iopscience.iop.org/1742-5468/2010/10/P10005)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

A hypergraph model of social tagging networks

Zi-Ke Zhang[1] **and Chuang Liu**[1,2,3]

[1] Department of Physics, University of Fribourg, Fribourg CH-1700, Switzerland
[2] School of Business, East China University of Science and Technology, Shanghai 200237, People's Republic of China
[3] Engineering Research Center of Process Systems Engineering (Ministry of Education), East China University of Science and Technology, Shanghai 200237, People's Republic of China
E-mail: zhangzike@gmail.com and liuchuang@mail.ecust.edu.cn

Received 24 July 2010
Accepted 14 September 2010
Published 7 October 2010

Online at stacks.iop.org/JSTAT/2010/P10005
doi:10.1088/1742-5468/2010/10/P10005

Abstract. The past few years have witnessed the great success of a new family of paradigms, so-called folksonomy, which allows users to freely associate tags with resources and efficiently manage them. In order to uncover the underlying structures and user behaviors in folksonomy, in this paper, we propose an evolutionary hypergraph model for explaining the emerging statistical properties. The present model introduces a novel mechanism that can not only assign tags to resources, but also retrieve resources via collaborative tags. We then compare the model with a real-world data set: *Del.icio.us*. Indeed, the present model shows considerable agreement with the empirical data in the following aspects: power-law hyperdegree distributions, negative correlation between clustering coefficients and hyperdegrees, and small average distances. Furthermore, the model indicates that most tagging behaviors are motivated by labeling tags on resources, and the tag plays a significant role in effectively retrieving interesting resources and making acquaintances with congenial friends. The proposed model may shed some light on the in-depth understanding of the structure and function of folksonomy.

Keywords: growth processes, network dynamics, online dynamics

Contents

1. Introduction

Networks provide us with a powerful and versatile tool for recognizing and analyzing complex systems where nodes represent individuals, and links denote the relations between them. Recently, many efforts have been made to understand the structure, evolution and dynamics of complex networks [1]–[5]. The advent of Web 2.0 and its affiliated applications brought in a new form of user-centric paradigm which cannot be fully described by pre-existing models on unipartite or bipartite networks. One such example is the user-driven emerging phenomenon, *folksonomy*, which allows users to upload resources (bookmarks, photos, movies, etc) and freely assign them with user-defined words, so-called *tags*. Folksonomy requires no specific skills for users to participate in it, broadens the semantic relations among users and resources, and eventually achieves its immediate success in a few years. Currently, a large number of such applications can be found online, such as *Del.icio.us* [6], *Flickr* [7], *CiteULike* [8], etc. With the help of those platforms, users can not only store their own resources and manage them with collaborative tags, but also look into other users' collections to find what they might be interested in by simply keeping track of the baskets with tags. Unlike traditional information management methods where words (or indices) are normally pre-defined by experts or administrators, e.g. the library classification systems, a tagging system allows users to create arbitrary tags that even may not exist in dictionaries. Therefore, those user-defined tags can reflect user behaviors and preferences using which users can easily make acquaintance, collaborate and eventually form communities with others who have similar interests [9].

Up to now, a variety of research works have been done in realizing the structure and dynamic process of folksonomy. Golder *et al* studied the usage patterns of *collaborative tagging systems* and classified seven kinds of tag functions [10], which is very helpful for us in achieving a better understanding of both the user behaviors and tagging purposes. In addition, the keywords or PACS numbers based methods are put forward as revealing the underlying structure of co-authorship and citation networks [11, 12]. Furthermore, many efforts have been made to explain how folksonomy emerges. Cattuto *et al* [13] investigated the dynamics of an open-ended system with a memory based Yule–Simon model. The model considered the ageing effect of tags, as well as the frequency of tag

J. Stat. Mech. (2010) P10005

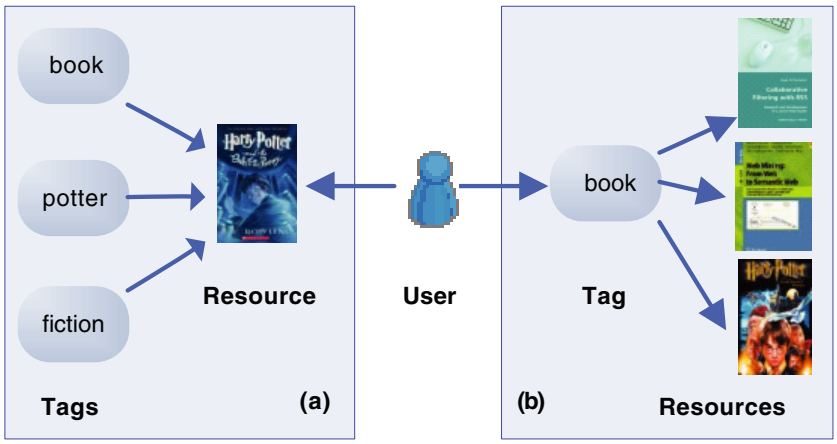

Figure 1. Illustration of two typical user tagging behaviors: (a) the user finds a resource (e.g. a book) via web surfing and annotates it with three tags for further use; (b) s/he collects one or some books by filtering out unrelated information with the tag 'book'.

occurrence. In [14], Lambiotte *et al* tried to model folksonomy in the form of tripartite graphs.

Recently, the *hypergraph* theory [15] allowed a hyperedge to connect an arbitrary number of vertices instead of two in regular graphs. Therefore, it provides us with a promising way to better understand a wide range of real systems. Up to now, there have been found applications in *personalized recommendation* [16]–[18], *population stratification* [19], *cellular networks* [20], etc. Besides, the definition is comparatively appropriate for uncovering underlying usage patterns and essential structures of folksonomies. Ghoshal *et al* [21] proposed a random hypergraph model to represent the ternary relationship where a hyperedge consists of one user, one resource and one tag, and reproduced many properties of folksonomy using the model. Zlatić *et al* [22] extensively defined a number of useful topological features based on hypergraph representation, which can be considered as a standard tool in understanding the structure of tagged networks.

In this paper, we propose a hypergraph model in order to illustrate the emergence of some statistical properties in folksonomy, including degree distribution, clustering coefficients and average distance between nodes. In this model, we consider two typical user tagging behaviors: (i) a user might be aware of a resource via web surfing or word-of-mouth propagation, and then save it as his/her own favorite collection and annotate it with tags of related topics for efficient management and retrieval; (ii) s/he might firstly pick up one or several compound tags, and then choose one possible resource from the retrieval results. Recently, a considerable amount of research has focused on the previous motivation [23, 24], while the latter is comparatively lacking attention. Actually, a tag is able to provide more relevant results according to its simple yet essential property of collaboration and semantics. Figure 1 shows those two different kinds of mechanisms.

In this model, users can manage resources with collaborative tags, and find resources using tags via serendipitous browsing. We then compare the model to one real-world data set, *Del.icio.us*, and find good agreement between them.

J. Stat. Mech. (2010) P10005

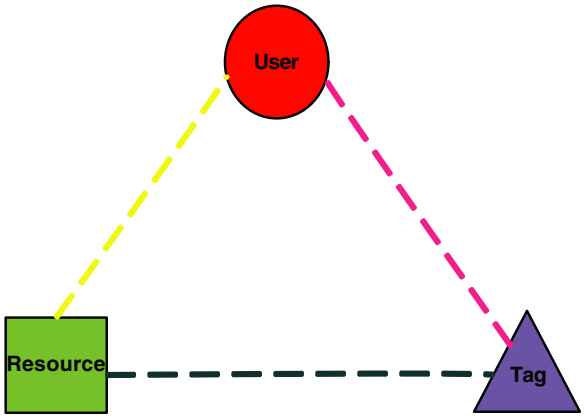

Figure 2. A hyperedge illustration of the basic unit in our network which was first introduced in [21]. There are vertices of three types in each hyperedge (represented as a triangle), depicted by one red circle, one green rectangle and one blue triangle, which respectively represent a user, a resource and a tag in folksonomy.

2. Modeling tripartite hypergraphs

We begin our study with some related definitions of tripartite hypergraphs that we will analyze. In this paper, we use the tripartite hypergraph representation given by [21], where a hyperedge simply consists of one user, one resource and one tag. Figure 2 gives a visual explication of such structure.

In a tripartite hypergraph, the network $\boldsymbol{G}$ can be briefly depicted by $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{H})$, where $\boldsymbol{V}$ denotes the vertices and $\boldsymbol{H}$ represents the set of hyperedges. $\boldsymbol{V} = \boldsymbol{U} \cup \boldsymbol{R} \cup \boldsymbol{T}$ where $\boldsymbol{U}$, $\boldsymbol{R}$ and $\boldsymbol{T}$ represent the set of users, resources and tags respectively, and $\boldsymbol{H} \subseteq \boldsymbol{U} \times \boldsymbol{R} \times \boldsymbol{T}$ is usually much smaller than the number of all the possible triangles. Correspondingly, the *Del.icio.us* data set that we collected has 15 009 users, 2431 190 resources and 325 120 distinct tags, which subsequently constitute 11 739 998 hyperedges.

The model

Consequently, we are mainly interested in the effect of tagging behaviors and the role of tags in networks. Therefore, we fix the distribution of user activities according to the empirical data. Thus, the model can be described as follows.

- At each time step, pick out a random user u according to the given distribution of user activities.
- For u, s/he can either choose a resource with probability p, or select an arbitrary tag with probability $1 - p$.
- If u is activated from the aspect of resource, s/he will randomly select an existing resource in the system with probability $1-p_1$ according to its popularity, or introduce a completely new resource with probability p_1. And then s/he will annotate it with a few tags. For simplicity, in this paper, we only consider that u will assign only

J. Stat. Mech. (2010) P10005

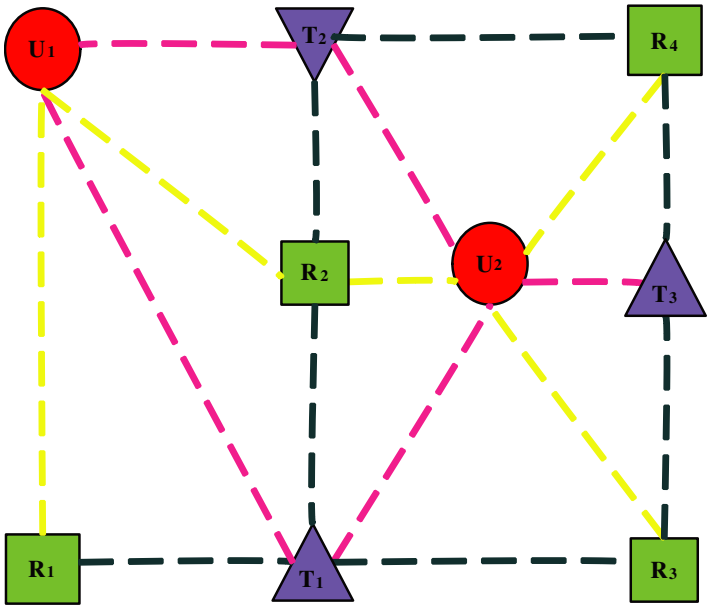

Figure 3. A descriptive hypergraph consists of two users, four resources and three tags. Take user U_2 and resource R_1 for example; the measurements are denoted as: (i) U_2 has participated in six hyperedges, which means its hyperdegree is 6; (ii) U_2 has directly connected to three resources and three tags, which suggests it possibly has $3 \times 3 = 9$ hyperedges maximally. Thus its clustering coefficient equals $6/9 \approx 0.667$, where its hyperdegree is 6; (iii) the shortest path from U_2 to R_1 is $U_2 - T_1 - R_1$, which indicates that the distance between U_2 and R_1 is 2.

one tag to the selected resource r. Thus, u could choose a tag t from his/her own vocabulary with probability p_2 according to how many times s/he has adopted it, or from the resource vocabulary with probability p_3 according to how many times it has been associated with the target resource, or introduce a new tag with probability $1-(p_2+p_3)$ from the initial tag pool of which the length, T_0, is large enough, if s/he does not find a suitable or personalized tag for describing r.

- If u decides to find a relevant resource from a specific topic, s/he will choose a random tag t based on its popularity, and then save one of the relevant resources according to in how many triangles they have appeared together with t.

In this model, a new hyperedge (u, r, t) is produced either from the perspective of resources or tags at each time step. When one tries to give a tag to a certain resource, s/he might choose a previous tag s/he used before, or pick up one tag recommended by the system. A new tag is added if no appropriate tag is available for describing that resource. Thus a tag-growth mechanism is considered in the present model. We then repeatedly run the model until enough hyperedges are obtained. Moreover, we simply assume that there is only one hyperedge emerging once the user is activated, which is a simplified case of real networks. However, such a simplified assumption could help us examine the effects of different tagging behaviors on the emergence of folksonomies. To evaluate our model, we measure the following quantities (figure 3 gives a detailed description of these quantities):

(i) *hyperdegree distribution*: defined as the proportion that each hyperdegree occupies, where hyperdegree is defined as the number of hyperedges that a regular node participates in;

(ii) *clustering coefficients*: defined as the proportion of real number of hyperedges to all the possible number of hyperedges that a regular node could have;

(iii) *average distance*: defined as the average shortest path length between two random nodes in the whole network.

Since we are mainly interested in how the tagging behaviors influence the emergence of folksonomies, we fix other parameters and investigate the effect of p. In the following analysis, we set $p_1 = 0.3$, $p_2 = p_3 = 0.45$ as constants.

2.1. Hyperdegree distribution

According to [21], hyperdegree is defined as how many triples a regular node takes part in. Thus we denote as $p_{(k_u)}$, $p_{(k_r)}$, $p_{(k_t)}$ the hyperdegree distributions of users, resources and tags, respectively. In terms of the model, $p_{(k_u)}$ is directly derived from the empirical data. Therefore, we mainly focus on the dynamics of $p_{(k_r)}$ and $p_{(k_t)}$. Firstly, we can write down the rate equation for the distribution of resources [1] (in order to avoid confusion of the time symbol, we use l to represent the time in the following descriptions):

$$\begin{aligned} p_{(k_r,l_i,l+1)} = p\{&p_1 p_{(k_r,l_i,l)} + (1-p_1)[(1-q_{(k_r,l)})p_{(k_r,l_i,l)} + (1-\delta_{k_r,1})q_{(k_r-1,l)}p_{(k_r-1,l_i,l)}]\} \\ &+ (1-p)\{(1-o_{(k_r,l)})p_{(k_r,l_i,l)} + (1-\delta_{k_r,1})o_{(k_r-1,l)}p_{(k_r-1,l_i,l)}\} + \frac{1}{l}\delta_{k_r,1}, \end{aligned} \tag{1}$$

where $p_{(k_r,l_i,l)}$ is denoted as the probability that in the time l a resource introduced at time l_i has a hyperdegree k_r, $q_{(k_r,l)}$ is the probability of picking up an uncollected resource with hyperdegree k_r for u according to r's popularity, $o_{(k_r,l)} = k_r/l$ is the probability of choosing a resource from a random tag t at time l, and δ is the Kronecker delta. The first brace shows the choice described in the model, where the first term is the probability of adding a new resource, the second term is the probability of selecting an existing resource which consists of addition two terms: (i) the probability of selecting a resource with hyperdegree k_r; (ii) the probability of not picking up a resource with hyperdegree $k_r - 1$. The second brace depicts the evolutionary process from the aspect of tags, in which the first term is the probability of selecting a resource with hyperdegree k_r and the second term is the probability of not picking up a resource with hyperdegree $k_r - 1$. The last δ term is the effect on the resources with hyperdegree $k_r = 1$ of introducing a new resource at time l. However, it is not easy to identify the distribution of each individual's absent resources; we consider approximatively that the distribution is in direct proportion to that of the system, that is,

$$q_{(k_r,l)} \approx \frac{k_r}{l}. \tag{2}$$

Integrating equations (1) and (2), as well as the stationary condition $p_{(k_r)} = \lim_{l\to\infty}(\sum_{l_i} p_{(k_r,l_i,l)})/l$, we have

$$p_{(k_r)} \approx \left(\frac{k_r - 1}{k_r + 1/(1-pp_1)}\right) p_{(k_r-1)}, \tag{3}$$

J. Stat. Mech. (2010) P10005

for $k_r > 1$. When $k_r = 1$, equation (1) can be simplified to

$$p_{(k_r=1)} = \frac{1}{2 - pp_1}. \tag{4}$$

Combining equations (3) and (4), we can recursively obtain the final solution:

$$p_{(k_r)} \approx a_1 \frac{\Gamma(k_r)\Gamma(1+a_1)}{\Gamma(k_r+1+a_1)} = a_1 B(k_r, 1+a_1), \tag{5}$$

where $a_1 = 1/(1-pp_1)$, Γ is the Gamma function and B is the beta function. Note that when k_r is large, equation (5) can be approximated as

$$p_{(k_r)} \approx a_1 \Gamma(1+a_1) k_r^{-(1+a_1)} \propto k^{-(1+a_1)}. \tag{6}$$

Analogously, we can also write down the tag hyperdegree distribution in the form of a rate equation:

$$\begin{aligned} p_{(k_t,l_i,l+1)} = {} & [(1-p) + p(p_2+p_3)][s_{(k_t-1,l)} p_{(k_t-1,l_i,l)}(1-\delta_{k_t,1}) \\ & + (1 - s_{(k_t,l)}) p_{(k_t,l_i,l)}] + p(1-p_2-p_3) p_{(k_t,l_i,l)} + \frac{1}{l}\delta_{k_t,1}, \end{aligned} \tag{7}$$

where $s_{(k_t,l)}$ is the probability of picking up a random tag with hyperdegree k_t at time l. According to the present model, there are four mechanisms that drive the growth of tags: (i) user u selects one tag from his/her own vocabulary with probability p_2; (ii) u chooses one word from the set of tags associated with the target resource with probability p_3; (iii) a new tag is introduced with probability $1-p_2-p_3$; (iv) u selects an interesting tag t from all the possible candidates and saves a resource that is relevant with t. Equation (7) exactly expresses the integrated effect on tag evolution of those mechanisms.

We make an assumption similar to equation (2), that the individual's tag hyperdegree distribution is in direct proportion to that of the system:

$$s_{(k_t,l)} \approx \frac{k_t}{l}. \tag{8}$$

We then follow the same processes as for equations (3) and (4); the solution will read

$$p_{(k_t)} \approx a_2 B(k_t, 1+a_2), \tag{9}$$

where $a_2 = 1/(1-p(1-p_2-p_3))$. Analogously to the case for equation (7), when k_t is large, equation (9) can be approximated as

$$p_{(k_t)} \approx a_2 \Gamma(1+a_2) k_t^{-(1+a_2)} \propto k_t^{-(1+a_2)}. \tag{10}$$

Equations (6) and (10) show that both the resource and tag hyperdegree distributions follow power laws when k_r and k_t are large. Additionally, we find that the scale-free property can also be found in small hyperdegrees with equations (5) and (9). Therefore, we will use equations (5) and (9) for further discussions. Figure 4 shows the simulation, analytical and empirical results of hyperdegree distributions in both the real and modeled networks. Figure 4(a) shows the empirical data of users' cumulative hyperdegree distributions, which follows a stretched exponential distribution [25, 26]. Figures 4(b) and (c) show good agreements among empirical observations and analytical

J. Stat. Mech. (2010) P10005

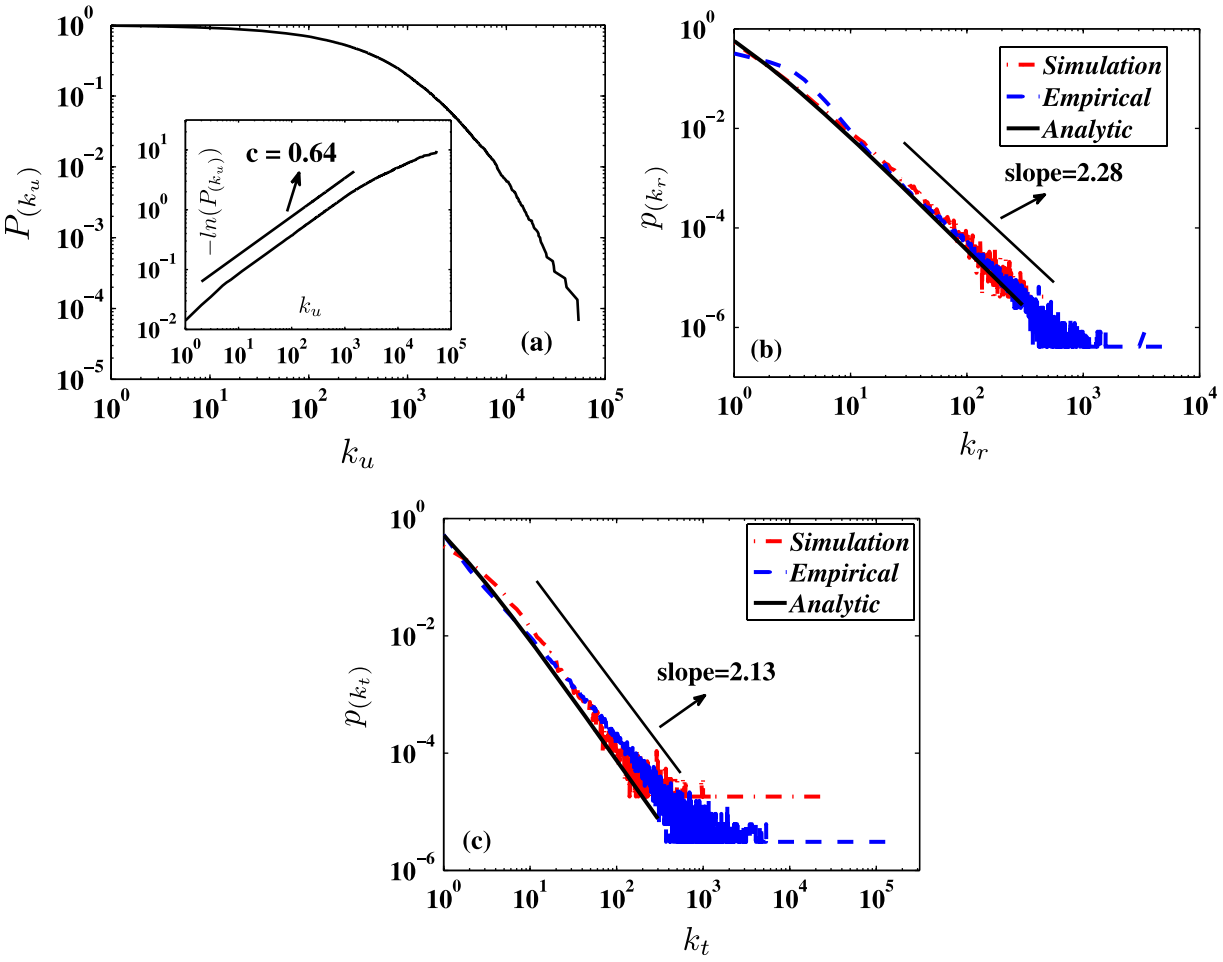

Figure 4. The hyperdegree distributions of nodes of three types: (a) the empirical cumulative hyperdegree distribution of users which follows a stretched exponential distribution $P(k_{\mathrm{u}}) \propto \exp^{-(k_{\mathrm{u}}/k_0)^c}$, where k_0 is a constant—the inset gives the fitting result for the exponent $c = 0.64$ according to the method used in [27]; (b) the empirical, simulation and analytical results for the resource hyperdegree distribution, following the power law $p_{(k_{\mathrm{r}})} \propto k_{\mathrm{r}}^{-\phi}$ with $\phi = 2.28$; (c) the empirical, simulation and analytical results for the tag hyperdegree distribution, following the power law $p_{(k_{\mathrm{t}})} \propto k_{\mathrm{t}}^{-\varphi}$ and $\varphi = 2.13$. The simulation and analytical results of (b) and (c) are obtained when $p = 0.8$.

results, while the inconsistency in figure 4(b) might be caused by our assumption that results in a comparatively large number resources with small hyperdegrees. Note that $p = 0.8$ indicates that most actions of the tagging are from the resource aspect in folksonomies.

In addition, we measure the effect on hyperdegree distribution of different values of p. In figure 5(a), the resource hyperdegree distribution is in good agreement only when p increases over 0.7, whereas the slope of the tag hyperdegree distribution does not change much with various values of p. This might be caused by two factors: (i) the evolution of folksonomy is driven primarily by assigning tags to the target resource, which is consistent

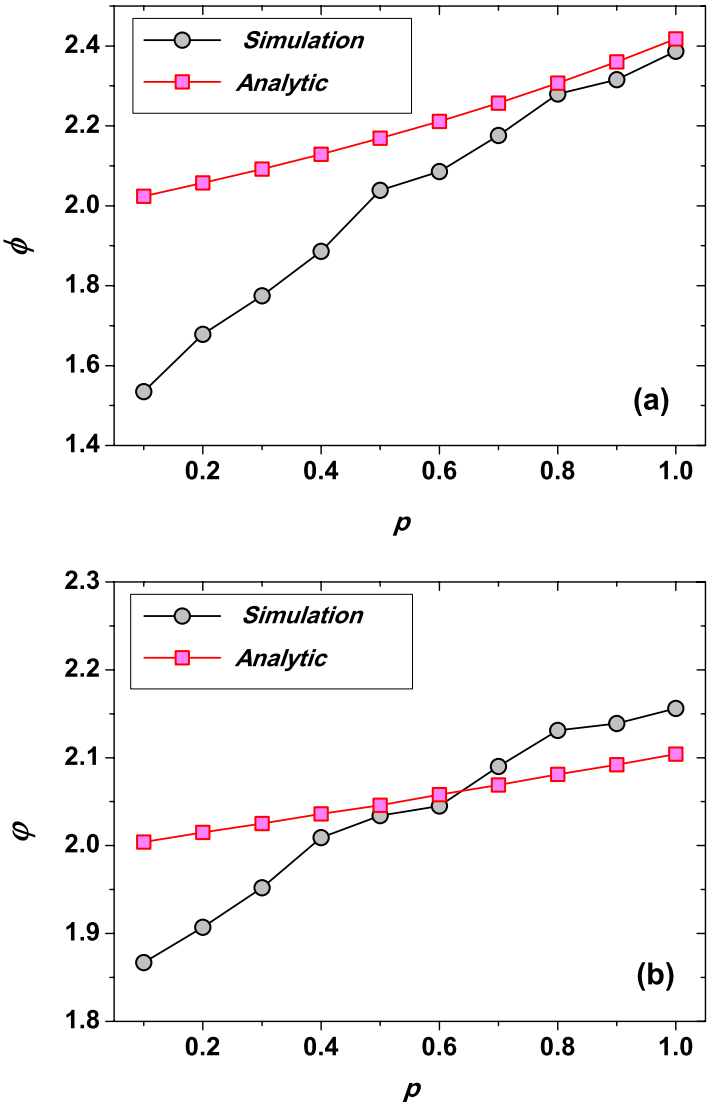

Figure 5. The slopes of hyperdegree distribution change according to different values of p for analytical and simulation results. (a) The variation of ϕ. (b) The variation of φ. Both of the distributions show scale-free properties for disparate values of p, that is, $p(k) \propto k^{-\alpha}$, where α refers to ϕ and φ in (a) and (b), respectively. ϕ and φ are measured by the least squares method (LSM).

with large values of p; (ii) when p is small, the fat tail of resources with small degree will remarkably affect the fitting result.

2.2. Clustering coefficients

Clustering in a network measures the likelihood that two neighbors of a given node are inclined to connect to each other. Watts and Strogatz [28] have introduced the *clustering coefficient* to measure the amount of clustering for a given node in normal unipartite networks. However, this definition is not fully compatible with the hypergraph case, since

J. Stat. Mech. (2010) P10005

a regular node connects two other different kinds of nodes. Thus, we adopt the definition of the user clustering coefficient given in [29][4]:

$$C_{\mathrm{u}} = \frac{k_{\mathrm{u}}}{R_{\mathrm{u}} \cdot T_{\mathrm{u}}}, \tag{11}$$

where k_{u} is the hyperdegree of user u, R_{u} is the number of resources that u collects and T_{u} is the number of tags that u possesses. The above definition measures the fraction of possible pairs present in the neighborhood of u. A larger C_{u} indicates that u has more similar topics of resources, which might also show that u has more concentrated on personalized or special topics, while smaller C_{u} might suggest that s/he has more diverse interests. Then the hyperdegree based clustering coefficient is averaged over all the users with the same hyperdegrees.

In order to compute C_{u}, we shall consider the evolutionary dynamics of T_{u}, the number of tags used by the selected user, as well as the dynamics of T, the current number of tags existing in the system. We can write the differential functions as

$$\begin{aligned}
\frac{\mathrm{d}_{T_{\mathrm{u}}}}{\mathrm{d}_l} &= \frac{k_{\mathrm{u}}}{L}\left[pp_3(1-p_1)\left(1-\frac{T_{\mathrm{u}}}{T}\right) + p(1-p_2-p_3)\left(1-\frac{T_{\mathrm{u}}}{T_0}\right) + (1-p)\left(1-\frac{T_{\mathrm{u}}}{T}\right)\right], \\
\frac{\mathrm{d}_T}{\mathrm{d}_l} &= p(1-p_2-p_3)\left(1-\frac{T}{T_0}\right),
\end{aligned} \tag{12}$$

where T_0 is the total number of tags that we initially set at the beginning of the model and L is the total number of designed simulation steps. Since we assume that only one tag is allowed to be assigned at each time step, the hyperdegrees of users and resources are degenerate to those for the bipartite case. Therefore, we get $k_{\mathrm{u}} = R_{\mathrm{u}}$. Thus, equation (11) can be rewritten as

$$C_{\mathrm{u}} = \frac{1}{T_{\mathrm{u}}}. \tag{13}$$

Unfortunately, it is not easy to get the explicit expression for equation (12). Instead, we find the numerical solution by combining equations (12) and (13). Figure 6(a) shows the good consistency among the empirical, simulation and numerical results.

Analogously, we can also write the dynamics of C_{r}:

$$\begin{aligned}
\frac{\mathrm{d}_{T_{\mathrm{r}}}}{\mathrm{d}_l} &= \frac{k_r}{l}p\left[p_2(1-p_1)\left(1-\frac{T_{\mathrm{r}}}{T}\right) + (1-p_2-p_3)\left(1-\frac{T_{\mathrm{r}}}{T_0}\right)\right], \\
\frac{\mathrm{d}_T}{\mathrm{d}_l} &= p(1-p_2-p_3)\left(1-\frac{T}{T_0}\right), \\
\frac{\mathrm{d}_{k_r}}{\mathrm{d}_l} &= \frac{k_r}{l}, \\
C_{\mathrm{r}} &= \frac{k_{\mathrm{r}}}{U_{\mathrm{r}} \cdot T_{\mathrm{r}}} = \frac{1}{T_{\mathrm{r}}},
\end{aligned} \tag{14}$$

[4] To evaluate the clustering coefficients, C_k, in hypergraphs, where k is the hyperdegree, Zlatić *et al* [22] proposed a metric based on counting the overlaps of a coordination number, z, for a given vertex, which gives a meaningful measurement of C_k. Unfortunately, it is not easy analytically to obtain the dynamics of C_k with this definition. We, therefore, as an alternative, adopt the definition of [29] in this paper, which is simple and easy for us to use to mathematically analyze the dynamics of C_k for tripartite hypergraphs.

J. Stat. Mech. (2010) P10005

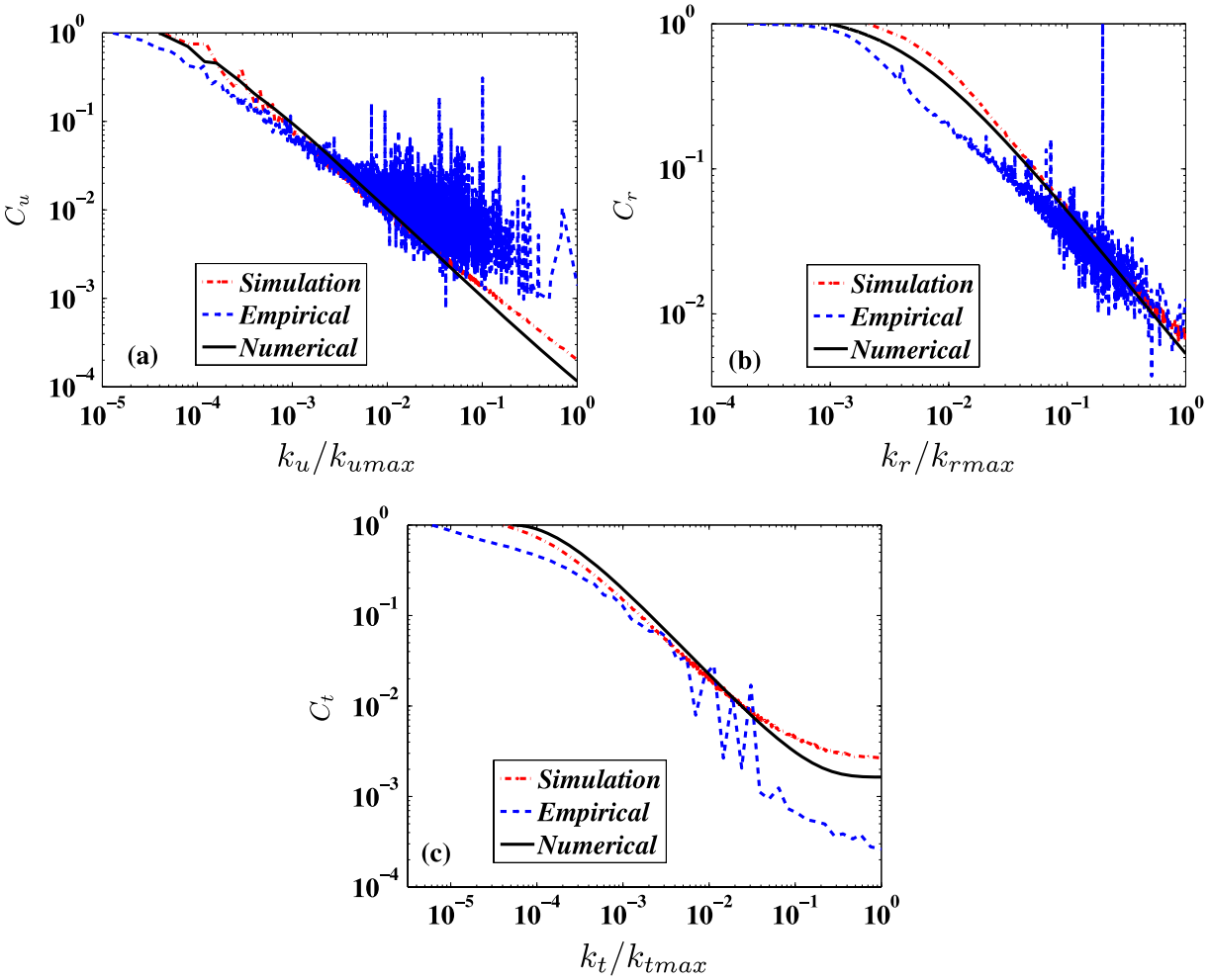

Figure 6. The clustering coefficients versus collapsed hyperdegrees. (a) The user clustering coefficient versus collapsed user hyperdegree; (b) the resource clustering coefficient versus collapsed resource hyperdegree; (c) the tag clustering coefficient versus collapsed tag hyperdegree. In (c), the empirical data set is shown in log bin form in order to alleviate the fluctuation resulting from insufficient data, interfering with the exhibiting of the statistical properties. All three plots are obtained with $p = 0.8$.

where k_{r} is the resource hyperdegree, T_{r} is the number of tags attached to resource r, and U_{r} is the number of users who have collected r. Figure 6(b) shows the numerical solution for equation (14), as well as the empirical and simulation results.

And the dynamics of C_{t} is as follows:

$$\frac{\mathrm{d}_{U_{\mathrm{t}}}}{\mathrm{d}_l} = \frac{k_t}{l}\left[(1-p)\left(1-\frac{U_{\mathrm{t}}}{U}\right) + p_3(1-p_1)\left(1-\frac{U_{\mathrm{t}}}{U}\right)\right],$$
$$\frac{\mathrm{d}_{k_t}}{\mathrm{d}_l} = \frac{k_t}{l},$$

$$\frac{\mathrm{d}_{R_\mathrm{t}}}{\mathrm{d}_l} = \frac{k_t}{l} p \left[p_2(1-p_1)\left(1-\frac{R_\mathrm{t}}{R}\right) + p_1(p_2+p_3) \right],$$
$$\frac{\mathrm{d}_R}{\mathrm{d}_l} = pp_1,$$
$$C_\mathrm{t} = \frac{k_\mathrm{t}}{U_\mathrm{t} \cdot R_\mathrm{t}}, \tag{15}$$

where k_t is tag hyperdegree, U is the number of users which is fixed in the model, U_t is the number of users who have used tag t, R is the number of resources existing in the system, and R_t is the number of resources labeled with t. Figure 6(c) shows the numerical solution for equation (15), as well as the empirical and simulation results. All three plots in figure 6 show negative correlations between clustering coefficients and hyperdegrees in both the real-world and modeled networks. This might indicate a hierarchical structure of tripartite hypergraphs [30], and suggest that users with larger hyperdegrees have more diverse interests, and vice versa.

2.3. Average distance

Another important quantity is the distance, D, between a random pair of nodes in a network. Hence, the average distance, $\langle D \rangle$, measures the efficiency of retrieving a target node in a network. Take a friendship network for example; $\langle D \rangle$ is given by counting the average shortest path length between a random user and another arbitrary user. Therefore, $\langle D \rangle$ assesses how easily yet effectively a user can make acquaintance with others in a given friendship network.

However, in the case of a tripartite hypergraph, there are three different regular nodes. Therefore, the shortest path length can be defined as the minimal number of hyperedges that must be traversed to go from vertex to vertex. Figure 7 shows $\langle D \rangle$ between any two kinds of vertices. Figures 7(a) and (b) show the average distances of the bipartite network and hypergraph structure of *Del.icio.us*, respectively. We can see the following. (i) Tags can significantly shorten $\langle D \rangle$ for any pair of nodes in comparison with the bipartite case. For example, $\langle D \rangle$ for the user–user pair is enhanced from 3.587 to 2.205, $\langle D \rangle$ for the user–resource pair is improved from 3.947 to 2.676, and the value of $\langle D \rangle$ for the resource–resource pair is shortened from 4.641 to 3.386. These considerable improvements might indicate that tags play an important role in *information retrieval*. (ii) In figure 7(b), the magnitude strictly follows the order $D_\mathrm{u} < D_\mathrm{r} < D_\mathrm{t}$ in both general and special cases. For example, we have $D_\mathrm{uu} < D_\mathrm{ur} < D_\mathrm{ut}$ for users, $D_\mathrm{ur} < D_\mathrm{rr} < D_\mathrm{rt}$ for resources, and $D_\mathrm{ut} < D_\mathrm{rt} < D_\mathrm{tt}$ for tags. A similar pattern of these orders might imply that *Del.icio.us* is a user-centric system and so we can more easily find any information through users than in other ways. Besides, the main purpose of tagging is to more efficiently and effectively manage resources, which retains coherence for the comparatively large values of p in section 2.2. Figure 7(c) reproduces such exciting phenomena with $p = 0.8$ in the model. Furthermore, we study the effect of different values of p on the distances. In figure 7(d), it is shown that the order does stay almost steady whatever the value of p changes to. Additionally, figure 7(d) also indicates that all the distances decrease monotonically with the lessening of p, which might suggest that the more often we use tags, the more effectively we can find target information.

J. Stat. Mech. (2010) P10005

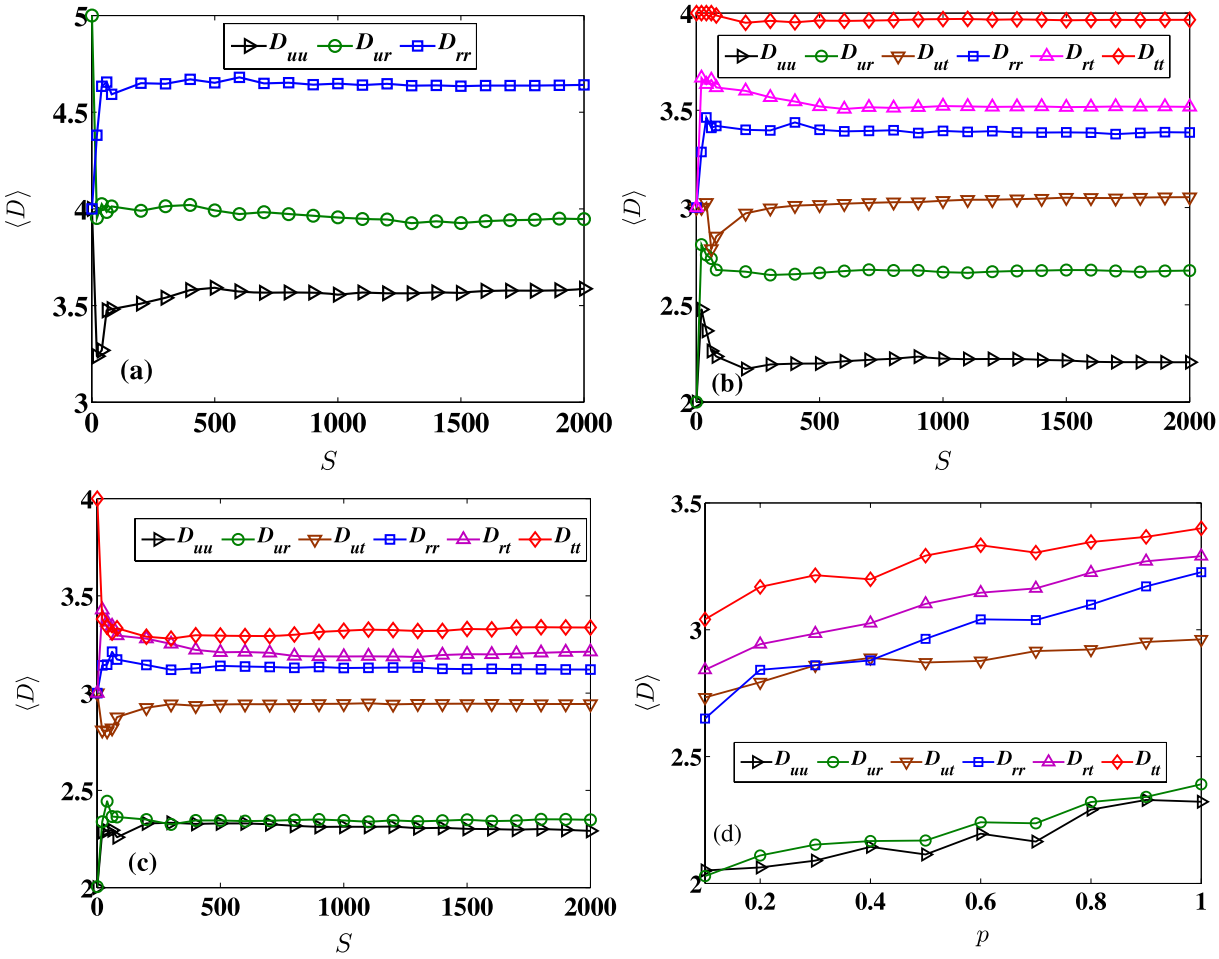

Figure 7. The average distances of bipartite and tripartite networks. Since the data set is huge, we calculate $\langle D \rangle$ by sampling randomly pairs of nodes until a stationary value is obtained. (a) The average user–user (D_{uu}), user–resource (D_{ur}) and resource–resource (D_{rr}) distances versus the number of samplings in the bipartite network, ignoring the tag information of *Del.icio.us*; (b) the average user–user, user–resource, user–tag (D_{ut}), resource–resource, resource–tag (D_{rt}) and tag–tag (D_{tt}) distances versus the number of samplings in the tripartite hypergraph of *Del.icio.us*; (c) the average user–user, user–resource, user–tag, resource–resource, resource–tag and tag–tag distances versus the number of samplings in the tripartite hypergraph produced by the present model—all the curves converge fast with just a small number of samplings, which indicates a small-world property in both bipartite and tripartite networks; (d) the stationary average distances change according to different values of p in the modeled network.

3. Conclusion and discussion

In this paper, we have proposed an evolutionary hypergraph model in order to study the dynamical properties of social tagging networks, so-called folksonomies. The present model assumes that there are two typical tagging behaviors based on the preferential

attachment mechanism: (i) assigning tags to users' favorite resources; (ii) saving resources that are relevant to interesting tags. The resulting tripartite hypergraph shows good agreement with a real-world network, *Del.icio.us*, on the following aspects: (i) the power-law hyperdegree distributions are generated for resources and tags, which indicates the heterogeneous topology; (ii) the average clustering coefficients decay with the increase of the hyperdegree, which may indicate hierarchical structure of tripartite hypergraphs; (iii) the average distances between vertices of the hypergraph are smaller than those in corresponding bipartite networks without tags; (iv) the relatively small average distance indicates a small-world property, which facilitates the *serendipitous discovery* of interesting contents and congenial companions; (v) all the above properties are found in relatively high consistency with a comparatively large value of $p = 0.8$, which suggests that the majority of actions are motivated by the first tagging behavior. Consequently, this model quantitatively reveals the accessorial yet significant role that tags play in folksonomies.

However, despite the good agreements in reproducing several features with real data, it is not easy to fully uncover the mechanisms dominating the emergence of folksonomy. This paper only provides a starting point for understanding the underlying motivations in facilitating a variety of intricate properties in such new paradigms. The present model considers that only one hyperedge is allowed to emerge at each time step, which is a moderately simplified version of real systems. In addition, users in different systems may have different tagging behaviors and the model should be improved to uncover the underlying mechanisms in other folksonomies. The tag co-occurrence [13, 29] and social cognitive imitation mechanisms [31] can be taken into account in order to improve the proposed model.

Acknowledgments

We acknowledge Dong Wei for providing us with the data set, and Jian-Guo Liu, Linyuan Lü and Chi-Ho Yeung for helpful discussions and suggestions. This work was partially supported by the Swiss National Science Foundation (Project 200020-121848). ZKZ acknowledges the National Natural Science Foundation of China under the grants nos 60973069 and 90924011. CL and ZKZ acknowledge the Scholarship Program supported by China Scholarship Council (CSC Program).

References

[1] Albert R and Barabási A-L, *Statistical mechanics of complex networks*, 2002 *Rev. Mod. Phys.* **74** 47
[2] Dorogovtsev S N and Mendes J F F, *Evolution of networks*, 2002 *Adv. Phys.* **51** 1079
[3] Newman M E J, *The structure and function of complex networks*, 2003 *SIAM Rev.* **45** 167
[4] Boccaletti S, Latora V, Moreno Y, Chavez M and Huang D-U, *Complex networks: structure and dynamics*, 2006 *Phys. Rep.* **424** 175
[5] Costa L da F, Rodrigues F A, Travieso G and Boas P R U, *Characterization of complex networks: a survey of measurements*, 2007 *Adv. Phys.* **56** 167
[6] http://del.icio.us.
[7] http://www.flickr.com.
[8] http://www.citeulike.com.
[9] Sen S, Lam S K, Rashid A M, Cosley D, Frankowski D, Osterhouse J, Harper F M and Riedl J, *Tagging, community, vocabulary, evolution*, 2006 *Proc. 20th Anniversary Conf. Computer Supported Cooperative Work* p 190

J. Stat. Mech. (2010) P10005

[10] Golder S A and Huberman B A, *Usage patterns of collaborative tagging systems*, 2006 *J. Inform. Sci.* **32** 198
[11] Palla G, Farkas I J, Pollner P, Deréyi I and Vicsek T, *Fundamental statistical features and self-similar properties of tagged networks*, 2008 *New J. Phys.* **10** 123026
[12] Zhang Z-K, Lü L, Liu J-G and Zhou T, *Empirical analysis on a keyword-based semantic system*, 2008 *Eur. Phys. J.* B **66** 557
[13] Cattuto C, Loreto V and Pietronero L, *Semiotic dynamics and collaborative tagging*, 2007 *Proc. Nat. Acad. Sci.* **104** 1461
[14] Lambiotte R and Ausloos M, *Collaborative tagging as a tripartite network*, 2006 *Lect. Not. Comput. Sci.* **3993** 1114
[15] Karypis G, Aggarwal R, Kumar V and Shekhar S, *Multilevel hypergraph partitioning: application in VLSI domain*, 1997 *Proc. 34th Annual Conf. Design Auto.* p 526
[16] Blattner M, *B-Rank: a top N recommendation algorithm*, 2009 arXiv:0908.2741
[17] Zhang Z-K, Zhou T and Zhang Y-C, *Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs*, 2010 *Physica* A **389** 179
[18] Shang M-S, Zhang Z-K, Zhou T and Zhang Y-C, *Collaborative filtering with diffusion-based similarity on tripartite graphs*, 2010 *Physica* A **389** 1259
[19] Vázquez A, *Population stratification using a statistical model on hypergraphs*, 2008 *Phys. Rev.* E **77** 066106
[20] Klamt S, Haus U-U and Theis F, *Hypergraphs and cellular networks*, 2009 *PLoS Comput. Biol.* **5** e1000385
[21] Ghoshal G, Zlatić V, Caldarelli G and Newman M E J, *Random hypergraphs and their applications*, 2009 *Phys. Rev.* E **79** 066118
[22] Zlatić V, Ghoshal G and Caldarelli G, *Hypergraph topological quantities for tagged social networks*, 2009 *Phys. Rev.* E **80** 036118
[23] Cattuto C, Barrat A, Baldassarri A, Schehr G and Loreto V, *Collective dynamics of social annotation*, 2007 *Proc. Nat. Acad. Sci.* **106** 10511
[24] Halpin H, Robu V and Shepherd H, *The complex dynamics of collaborative tagging*, 2009 *Proc. 16th. Conf. WWW* p 220
[25] Laherrère J and Sornette D, *Stretched exponential distributions in nature and economy: 'fat tails' with characteristic scales*, 1998 *Eur. Phys. J.* B **2** 525
[26] Shang M-S, Lü L, Zhang Y-C and Zhou T, *Empirical analysis of web-based user–object bipartite networks*, 2009 *Europhys. Lett.* **90** 48006
[27] Zhang P-P, Chen K, He Y, Zhou T, Su B-B, Jin Y-D, Chang H, Zhou Y-P, Sun L-C, Wang B-H and He D-R, *Model and empirical study on some collaboration networks*, 2006 *Physica* A **360** 599
[28] Watts D J and Strogatz S, *Collective dynamics of 'small-world' networks*, 1998 *Nature* **393** 440
[29] Cattuto C, Schmitz C, Baldassarri A, Servedio V D P, Loreto V, Hotho A, Grahl M and Stumme G, *Network properties of folksonomies*, 2007 *AI Commun.* **20** 245
[30] Ravasz E and Barabási A-L, *Hierarchical organization of modularity in complex networks*, 2003 *Phys. Rev.* E **67** 026112
[31] Dellschaft K and Staab S, *An epistemic dynamic model for tagging systems*, 2008 *Proc. 19th ACM Conf. Hypertext Hypermedia* p 71

J. Stat. Mech. (2010) P10005