2009 Fifth International Joint Conference on INC, IMS and IDC

A Global Social Graph as a Hybrid Hypergraph

Joonhyun Bae and Sangwook Kim
Electrical Engineering and Computer Science
Kyungpook National University, Daegu, Korea
{jhbae, swkim}@cs.knu.ac.kr

Abstract—**The emergence of large-scale online social networks has attracted much interest in recent years, and the structure of online social networks has been rigorously studied in the last few years. However, to date, only selected silos of fragmented online social networks have been investigated. This is due to the lack of information about the people who participate in several online social networks. However, recently, the evolution of portable social environment makes it possible to mine the global social graph of collective online social networks. Hence, in this paper, we study the structure of the global social graph as a *hybrid hypergraph*, where the hyperedges connect a user's multiple identities distributed over several local social graphs. Based on our empirical study using the Social Graph API, we show that (1) the population of websites and the degree distribution of hyperedges follow the power-law, and that (2) the existence of *connectors*, who participate in several online social networks, ensures that the global social graph is not fragmented but interconnected. We believe that these findings can shed a new light on the design of future systems based on decentralized social networks, with a bird's-eye view of the global social graph.**

Keywords*–*Online Social Network, Portable Social Environment, Global Social Graph, Hybrid Hypergraph

I. INTRODUCTION

The structure and evolution of online social networks have been rigorously studied in the last few years [1-7]. However, to our knowledge, all the prior studies have been restricted to some selected silos of fragmented online social networks, i.e., social networks formed in isolated websites, e.g., Twitter, LiveJournal, or Flickr. This is due to the lack of information about the people who participate in several online social networks. However, more recently, much attention has been paid to the portable social environment [8-12], making the information about the relationships among people to be a community asset. These moves toward a public, portable, and decentralized social networks lead us to the model of the global social graph, breaking the boundaries between fragmented online social networks.

In this paper, we propose a novel model for the global social graph of collective online social networks. We first define a *local social graph* as a social network formed by the users of a website. The vertices and edges in a local social graph are users' identifiers and their relationships in a site. Since a person can be appeared in several local social graphs with multiple identities, a vertex in a *global social graph* should be a set of local vertices distributed over several local social graphs. Hence, we present a *hybrid hypergraph model*, where the hyperedges connect a person's multiple identities. Hence, a global social graph as a hybrid hypergraph can be viewed as a unified graph of local social graphs interconnected by the hyperedges.

The hybrid hypergraph model can be empirically studied by the existing technologies, e.g. XFN[1] or FOAF[2], open standards for describing the connections among people. The Social Graph API[3], provided by the Google, also makes it possible to mine the information about various local social graphs and the hyperedges. Using the Social Graph API, we have mined about 1.09 million local vertices and 25.8 million local edges. The dataset we collected contains 60 websites and 0.87 million hyperedges. Our empirical study with this dataset shows that the global social graph is not fragmented but interconnected by the *connectors*, who have multiple identities in several websites. Moreover, in this corpus, we have found that the population of websites follows the power-law with the scaling parameter $\alpha \sim 1.77$. Also, to our surprise, the degree distribution of hyperedges follows the power-law with the power-law exponent $\beta \sim 3.29$. It implies that there exist the connectors who participate in several local social graphs, and the existence of connectors ensures that a global social graph is not fragmented but interconnected. We believe that these findings can shed a new light on the design of future systems based on decentralized social networks, with a bird's-eye view of the global social graph.

In this paper we also introduce some algorithms for the global social graph. We first provide two transformation mechanisms, what we call *canonicalization* and *normalization*, to convert a hybrid hypergraph into a *normalized weighted graph* via a *canonical multigraph*, where the hyperedges are treated as the vertices representing people's globally unique identities. We show that the time and space complexity of these algorithms is $O(|E|)$, where E is a union of local edges. Then we provide an efficient algorithm to find covert links, i.e., hidden local relationships which are able to be induced by the links in other local social graphs. The time complexity of finding covert links in a new website for a person is also proven to be $O(|h_s| \cdot |h_t|)$, where h_s and h_t are the hyperedges of source and target local vertices.

The rest of this paper is organized as follows. In Section 2, we review related work with regard to online social networks. In Section 3, we describe our models for the global social graph of collective online social networks. In Section 4, we

[1] *http://gmpg.org/xfn*
[2] *http://www.foaf-project.org*
[3] *http://code.google.com/apis/socialgraph*

978-0-7695-3769-6/09 $26.00 © 2009 IEEE
DOI 10.1109/NCM.2009.20

present the result of our empirical study using the Social Graph API. In Section 5, we introduce some algorithms for our proposed model. Finally, we conclude this paper in Section 6.

II. Related Work

Watts and Strogatz [13] discovered the small-world, where the short paths of acquaintances, linking all the people in the world, could be found. Barabási and Albert [14] showed that many complex networks, including social networks, are scale-free networks, where the degree distribution follows the power-law.

The structure and evolution of online social networks have been rigorously studied in the last few years. Kumar et al. [1] evaluated the growth of large-scale online social networks. They exposed three segmentations of social networks: singletons, isolated communities, and a giant connected component. Ahn et al. [2] compared the structures of three online social networks. They showed that the pattern of degree correlations is similar to real-life social networks. Mislove et al. [3] presented valuable findings on the properties of large-scale online social networks. Their measurements revealed that online social networks are the *small-world* and *scale-free* networks at the same time. In [4], Mislove et al. also examined the growth of the Flickr's online social network. They found that the links tend to be created by users who already have many links, users tend to respond to incoming links by creating links back to the source, and users tend to link to other users who are already close in the network. Cha et al. [5] investigated the way how information disseminates through social links in online social networks. They showed that social cascades are an important factor in the dissemination of content. Leskovec et al. [6] analyzed the edge-by-edge evolution of large online social networks and showed that most new edges span very short distances, typically closing triangles, and they proposed a model of network evolution, incorporating node arrivals, edge initiation, and edge destination selection processes. Hu and Wang [7] also studied the structural evolution of a large online virtual community. Their work found that the scale growth of online social networks shows non-trivial S-shape.

However, note that these studies have been done with some selected silos of fragmented online social networks. So, it is possible to say that we have seen only a part of the whole online social networks, although a person can participate in several online social networks. This is due to the lack of information about the links indicating a person's multiple identities in several local social graphs.

However, more recently, the circumstances are changing with the evolution of portable social networks. Fitzpatrick [8] pioneered the concept of the Social Graph as a community asset. Moreover, to our surprise, he realized his own thoughts on the Social Graph by providing the Social Graph API. For this kind of innovation, Berners-Lee [9] said that the *Net* (i.e., the Internet) links computers, the *Web* (i.e., the World Wide Web) links documents, and the *Graph* (i.e., to say, the Giant Global Graph) connects *things*. Ramakrishnan and Tomkins [10] proposed the PeopleWeb, which is formed by the users and their interactions with increasingly rich content. They prospected that the emergence of two new capabilities on the Web, a global object model and a portable social environment, would radically transform the way people interact and discover information at online. Heyman [11] showed that several existing technologies could be part of the solution to make the social data to be portable. Breslin and Decker [12] proposed a possible solution to build semantic social networking into the fabric of the next-generation Internet, interconnecting both the content and the people in meaningful ways. They advocated that the Semantic Web can provide the interoperability among social networking sites.

These moves toward a public, portable, and decentralized social networks lead us to the model of the global social graph, breaking the boundaries between fragmented online social networks.

III. Models

In this section we describe various graph models for the global social graph. We begin by the definition of local social graph, hybrid hypergraph, canonical multigraph, and normalized weighted graph. Then we present a simplified example to build a global social graph.

A. Local Social Graph

Let $G_i=(V_i, E_i)$ be a *local social graph* of vertices V_i and edges E_i in a site i. The elements of V_i are users' identifiers, and those of E_i are their relationships in a site i. For example, when two people, $V_i=\{romeo, juliet\}$, are the members of $i=twitter.com$, they may have mutual relationship of *following* and *follower*, creating two distinguished edges, $E_i=\{(romeo, juliet), (juliet, romeo)\}$.

Since there have been various social networking sites, *romeo* and *juliet* might have been connected in the other sites. In other word, there exist various local social graphs, $G_1, G_2, \ldots, G_n$, in the world. The problem of modeling these fragmented local social graphs as a *global social graph*, $G=G_1 \otimes G_2 \otimes \ldots \otimes G_n$, is the challenge of this paper.

B. Hybrid Hypergraph

A *hypergraph* is a generalization of a graph, where *hyperedges* can connect any number of vertices. Let $G_H=(V, E, H)$ be a *hybrid hypergraph* of vertices V, edges E, and hyperedges H. Here V is a union of local vertices, $V=V_1 \cup V_2 \cup \ldots \cup V_n$, and E is a union of local edges $E=E_1 \cup E_2 \cup \ldots \cup E_n$. In addition, H is a set of hyperedges, where the element of H is a subset of local vertices, i.e., $H=\{ h \mid h=\{v_1, v_2, \ldots, v_m\}, v_i \in V\}$.

In our hybrid hypergraph model, hyperedges connect a person's multiple identifiers distributed over diverse local social graphs. For example, if *romeo* has joined three sites, $i_1=twitter.com$, $i_2=livejournal.com$, and $i_3=flickr.com$, a hyperedge $h=\{i_1.romeo, i_2.romeo, i_3.ilovejuliet\}$ connects these multiple identifiers, and it represents a unique identity of *romeo* globally. Note that he has a different identifier, *ilovejuliet,* in *flickr.com.* Therefore we should endow a hyperedge with an identifier which has a globally unique value in $H(G_H)$.

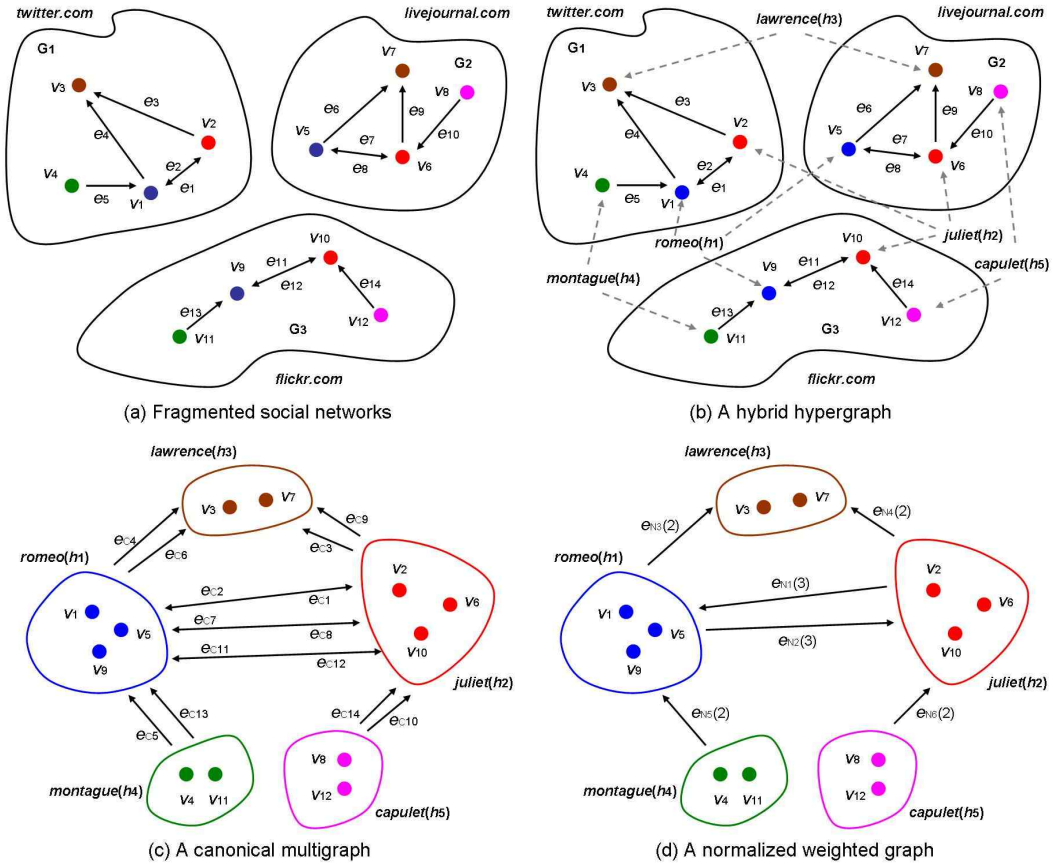


Figure 1. An illustrative example of building a global social graph. (a) is consist of three fragmented social networks, and (b) is a hybrid hypergraph, where the local social graphs are interconnected by 5 hyperedges. (c) is a canonical multigraph transformed by the hybrid hypergraph shown in (b). Finally, (d) is a normalized weighted graph, where the multiedges are combined into weighted edges.

C. *Canonical Multigraph*

A *multigraph* is a graph which is allowed to have *multiedges*, i.e., edges that have two identical sources and targets. Let $G_C=(H, E_C)$ be a *canonical multigraph* of hyperedges H and multiedges E_C. Here H is a set of hyperedges in G_H, and E_C is a set of multiedges which are transformed by edges in $E(G_H)$.

In our canonical multigraph model, hyperedges should be treated as vertices, and multiedges connect two hyperedges. Note that a hyperedge h_i in G_C represents an individual's unique identity in a global social graph. For example, if there exist two hyperedges, h_1 of *romeo* and h_2 of *juliet*, and two local edges, (i_1.*romeo, i*$_1$.*juliet*) and (i_2.*romeo, i*$_2$.*juliet*), local edges can be transformed into two multiedges, E_C= $\{(h_1, h_2, i_1), (h_1, h_2, i_2)\}$. Note that an element of E_C is *labeled*(or *annotated*) with the identifier of local social graph to discriminate the one from the other.

D. *Normalized Weighted Graph*

Let $G_N=(H, E_N)$ be a *normalized weighted graph* of hyperedges H and weighted edges E_N. Here H is also a set of hyperedges in G_H, but E_N is a set of normalized edges, where the multiedges in $E_C(G_C)$ which have the same endpoints are combined into an edge weighted by the number of links connecting two hyperedges.

In our normalized weighted model, edges can show the strength of the relationships between two people. To say, the heavier weight an edge has, the closer their relationship is. For example, suppose that two hyperedges, h_1 of *romeo* and h_2 of *juliet*, are connected by two bidirectional weighted edges, $E_N=\{(h_1, h_2, 3), (h_2, h_1, 3)\}$, where the cardinality of h_1 and h_2 are the same as the weight of their edges, i.e., $|h_1|=|h_2|=3$. In such a case, we can tell the strong relationship of *romeo* and *juliet*, since they are connected with each other in every site they have joined.

E. *Building a Global Social Graph*

As an illustrative example, let us consider three local social graphs of G_1=*twitter.com*, G_2=*livejournal.com*, and G_3= *flickr.com*. Figure 1(a) shows the structure of these exemplified social graphs which are formed by five individuals, *romeo*, *juliet*, *lawrence*, *montague*, and *capulet*. In Figure 1(a), a small circle is a local vertex, and a straight line is a local edge. As we can see here, there are 12 local vertices, i.e., $|V|=|V_1|+|V_2|+|V_3|=$ 4+4+4=12, and 14 local edges, i.e., $|E|=|E_1|+|E_2|+|E_3|=$ 5+5+4=14.

TABLE I. THE CHARACTERISTICS OF SOCIAL GRAPHS

Social Graphs		$\|V\|$	$\|E\|$	CC	α_{in}
G_1	*livejournal.com*	575,723	15,324,304	.10028	1.38
G_2	*flickr.com*	167,442	6,617,901	.10691	1.33
G_3	*twitter.com*	78,482	2,614,327	.13331	1.41
G_4	*digg.com*	21,273	437,303	.10303	1.47
G_5	*pownce.com*	24,229	321,000	.28812	1.62
G_6	*friendfeed.com*	12,095	167,913	.10023	1.67
G_7	*vox.com*	15,357	95,561	.06908	1.76
G_8	*zooomr.com*	10,579	93,263	.22137	1.90
G_9	*jaiku.com*	7,823	82,131	.26437	1.54
G_{10}	*last.fm*	12,858	19,334	.05106	2.11
$G_H=(V,E,H)=$ $G_1 \otimes G_2 \otimes \ldots \otimes G_{60}$		1,094,246	25,791,056	-	-
$G_C(G_H)=(H, E_C)$		871,393	25,791,056	-	-
$G_N(G_H)=(H, E_N)$		871,393	25,695,202	.08570	1.39

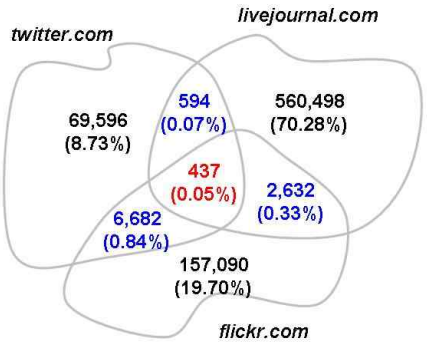


Figure 2. The existence of connectors between three large-scale local social graphs.

The basic problem of these models is to find hyperedges from those fragmented social networks. Although this problem is not a trivial one in real world, we assume that there are explicit links to local vertices from all the individuals' unique identities, represented by dotted lines in Figure 1(b). Hence there are five hyperedges, i.e., $|H|=5$, where the cardinalities of hyperedges are $|h_1|=|h_2|=3$ and $|h_3|=|h_4|=|h_5|=2$ respectively.

Now we have a hybrid hypergraph $G_H=(V, E, H)$, a canonical multigraph $G_C=(H, E_C)$ can be obtained by transforming $E(G_H)$ into $E_C(G_C)$. For example, let us consider $e_3 \in E_1$ and $e_9 \in E_2$ in Figure 1(b). Since the endpoints of $e_3=(v_2, v_3)$ are the elements of two different hyperedges, h_2 and h_3, respectively, it is transformed into an edge $e_{C3}=(h_2, h_3, G_1)$, annotated by G_1. With the same manner, $e_9=(v_6, v_7)$ is transformed into $e_{C9}=(h_2, h_3, G_2)$, annotated by G_2. The structure of the canonical multigraph G_C transformed by G_H is shown in Figure 1(c).

In the end, a normalized weighted graph $G_N=(H, E_N)$ can be obtained by combining all the multiedges in G_C which have the same endpoints into an edge weighted by the number of combined edges. For example, two annotated edges, $e_{C3}=(h_2, h_3, G_1)$ and $e_{C9}=(h_2, h_3, G_2)$, have the same endpoints both in a source and in a target. Therefore they can be combined into a normalized edge $e_{N4}=(h_2, h_3, 2)$, weighted by the number of links, i.e., 2. The final structure of normalized weighted graph is shown in Figure 1(d).

IV. MEASUREMENTS

In this section we present the result of our empirical study using the Social Graph API. What we have found is that the population of websites and the degree distribution of hyperedges follow the power-law. Here we show that the global social graph of collective online social networks are not fragmented but interconnected.

A. Mining the Global Social Graph

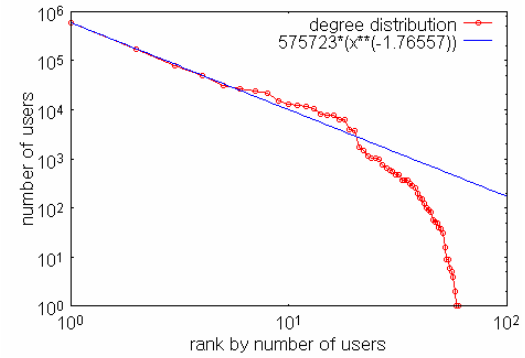
The primary challenge of our empirical study is to find and collect an appropriate dataset, where the social networks are not fragmented but interconnected by the hyperedges. To our knowledge, there has been no prior dataset including the hyperedges connecting a person's multiple identities from diverse online social networks. Here we present a detailed description of the presence of the hyperedges and describe the method of crawling a connected component using the Social Graph API.

XFN (XHTML Friends Network) is a simple way to represent human relationships using hyperlinks. It enables web authors to indicate their relationships to the people in their blogrolls by adding a ***rel*** attribute to the anchor tags. And the profile of XFN 1.1 includes a *me* type, a link to a self-identity at a different URL. For example, a hyperlink, <**a** ***href***="*http://brad.livejou-rnal.com*" ***rel***="*me*">, implies the author of this web page has an identity *brad* in a website *livejournal.com*. Thus it is easy and useful when pointing to various profiles on diverse social network sites or to several blogs maintained by the same person.

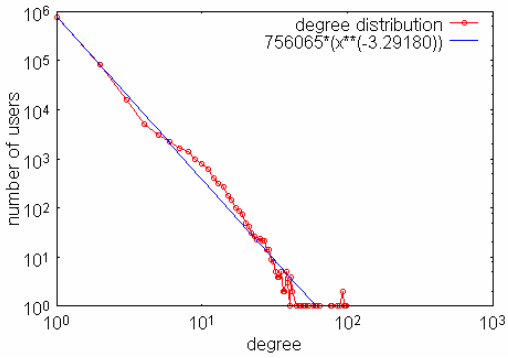
As is often the case with a large-scale link mining, crawling an entire connected component is not feasible. Mining the hyperedges widely dispersed over the Web, moreover, seems to be impossible at a first glance. Fortunately, the Google provides the Social Graph API. They are doing a good job of mining huge *me* links from the Web, and, made the corpus to be a community asset. Owing to the Social Graph API, what is to be done is as simple as crawling a global social graph, collecting the information about the hyperedges as well as the local edges.

The Social Graph API makes the information of public connections between people easily available at hands based on the open standards, XFN and FOAF. It returns the URLs of web pages and publicly declared connections among them. The *lookup* method lists all the local edges as well as the hyperedges from a given node. However, we use a special form of URL which is generated by the *Social Graph Node Mapper*[4], because it is difficult to identify the local graphs and personal identities from a generalized URL. For example, a node of a user, who has an identifier brad in a local graph livejournal.com would map to: "**sgn**://*livejour-nal.com* /?**ident**=*brad* ".

[4] *http://code.google.com/p/google-sgnodemapper*



(a) The population of websites



(b) The degree distribution of hyperedges

Figure 3. The population of websites and the degree distribution of hyperedges. (a) shows that the population of websites is governed by the power-law with the power-law exponent α ~1.77. (b) shows that the degree distribution of hyperedges follows the power-law with the scaling parameter β ~3.29.

Our crawler conducted a breadth first search starting from the Brad's identifier. Since the rate of sampling in this snowball-like method is known to produce a biased sample of nodes [15], our crawler follows all the forward links from a node. The dataset we gathered by this method using the Social Graph API contains 60 local graphs and 25.8 million local edges. The number of local vertices is about 1.09 million, and the number of hyperedges reaches 0.87 million. Hence, we can say that there exist 0.22 million duplicate identities. In this corpus, we have witnessed the existence of *connectors*, who have multiple identities in several sites, though many people are *dedicated* to a single site. Moreover, we have seen the existence of *navigators*, who have a lot of identities in many sites. Table I summarizes the dataset we collected for our empirical study.

B. Characterizing the Social Graph

In our corpus, we first examine two characteristic patterns of the clustering coefficient at microscopic level and the power-law exponent at macroscopic level.

The *clustering coefficient* [13] of a vertex in a graph quantifies how close a vertex and its neighbors are to being a *clique*, i.e., a complete graph. It is formally defined in a directed graph as the following:

$$C_i = \frac{\left|\{e_{jk}\}\right|}{k_i(k_i-1)}, v_j, v_k \in N_i, e_{jk} \in E \,, \tag{1}$$

where k_i is the degree of a vertex v_i, e_{jk} is an edge connecting two vertices from v_j to v_k, and N_i is a set of neighbors of v_i, i.e., $N_i = \{ v_j \mid e_{ij} \in E \}$. The clustering coefficient CC of a graph is also defined as:

$$CC = \frac{1}{N}\sum_{i=1}^{N} C_i \,, \tag{2}$$

Mathematically, a power-law is a special kind of relationship between two quantities. If a quantity x follows a power-law, it is drawn from a probability distribution:

$$p(x) \propto x^{-\gamma} \,, \tag{3}$$

where γ is a constant parameter of the distribution, which is known as the *power-law exponent* [14]. Typically, the degree distribution of complex networks is turned out to follow the power-law in the range $2 < \gamma < 3$.

The problem of fitting the power-law exponent can be done with the Maximum-Likelihood-Estimation method. In [16], Clauset et al. discussed that commonly used methods, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for power-law distributions, and they presented a principled statistical framework for quantifying power-law behavior in empirical data:

$$\gamma = 1 + n\left[\sum_{i=1}^{n} \ln\left(\frac{x_i}{x_{\min} - 1/2}\right)\right]^{-1} \,, \tag{4}$$

where x_i, i=1..n is the observed value of x such that $x_i \geq x_{\min}$.

Using those equations of (2) and (4), we measured the community structure by the clustering coefficient CC and the scale-freeness of in-degrees by the power-law exponent α_{in} . The figures measure in our corpus is presented in Table I, and we can see here the community structure and the scale-freeness are common in all the local social graphs and, especially, in a global social graph. Note that the resulting global graph parameters approximate the values measured in the largest local social graph, i.e., *livejournal.com*. It is measured that the clustering coefficient and power-law exponent of our are CC~0.0857 and α_{in} ~1.39 respectively. With this result, we can expect the global social graph can be characterized by the properties of local social graph both at microscopic and macroscopic levels.

C. Existence of the Connectors

We next investigate the population of websites and the degree distribution of hyperedges. As is often the case with the popularity of websites, the Zipf's law [17] may lead to the emergence of scaling in the population of websites. To verify this assumption, we plotted the number of users in a website by the ranking of the websites in our dataset as is shown in Figure 3(a). It is shown in this figure that the

ALGORITHM 1. CANONICALIZE.
INPUT: $G_H=(V, E, H)=G_1 \otimes G_2 \otimes \ldots \otimes G_n$
OUTPUT: $G_C=(H, E_C)$.
BEGIN.
$E_C \leftarrow \phi$.
for each $G_i=(V_i, E_i)$ **in** G_H.
for each $e_i=(v_s, v_t) \in E_i$, $v_s, v_t \in V_i$.
$h_s \leftarrow$ **find a hyperedge, where** $v_s \in h_s$.
$h_t \leftarrow$ **find a hyperedge, where** $v_t \in h_t$.
$E_C \leftarrow E_C \cup \{ e_C=(h_s, h_t, G_i) \}$.
END.

Figure 4. Pseudo-codes for *canonicalization*.

ALGORITHM 2. NORMALIZE.
INPUT: $G_C=(H, E_C)$.
OUTPUT: $G_N=(H, E_N)$.
BEGIN.
$E_N \leftarrow \phi$.
for each $e_C=(h_s, h_t, G_i) \in E_C$.
if $(e_N=(h_s, h_t, w) \in E_N$, **for any** $w)$ **then**
$e_N \leftarrow (h_s, h_t, w+1)$.
else
$E_N \leftarrow E_N \cup \{ e_N=(h_s, h_t, 1) \}$.
END.

Figure 5. Pseudo-codes for *normalization*.

population of websites is governed by the power-law with the scaling parameter $\alpha \sim 1.77$ as the following relationship:

$$p(w) = Cw^{-\alpha}, \quad (5)$$

where w is a ranking index, $p(w)$ is the fraction of users in a website with rank w, and C is a constant. This result implies that there are a few popular websites and a lot of small-sized websites.

Now let us consider the existence of the *connectors*, who participate in several local social graphs. Figure 2 shows the rate of connectors in three congested local social graphs, i.e., LiveJournal, Flickr, and Twitter. As we can see here, 437 people (0.05%) are appeared at all sites, and 9,908 connectors (1.24%) are the members of at least two sites. This result shows that there exist the connectors, though a lot of people are *dedicated* to a website. For the sake of characterizing the degree distribution of hyperedges, we draw a log-log plot as is shown in Figure 3(b). To our surprise, the degree distribution of hyperedges is turned out to follow the power-law with the power-law exponent $\beta \sim 3.29$ as the following relationship:

$$p(k) = Ck^{-\beta}, \quad (6)$$

where k is the degree of hyperedge, $p(k)$ is the fraction of hyperedges with k degree, and C is a constant. This result also implies that there exist a few active users, what we call *connectors*, and a lot of users who are dedicated to a single website.

To characterize the connectivity of websites, we measure the degree of connections among websites as the following:

$$C(G_H) = \frac{|E| - |H|}{|H|}, \quad (7)$$

where E is a union of local edges and H is a set of hyperedges. Here the connectivity $C(G_H)$ of a global social graph denote the rate of multiple edges from a set of hyperedges. If there is no multiple edge, the connectivity would be measured to be $C(G_H)=0$. In the other side, if all the users are appeared in every website, the connectivity would be measured to be $C(G_H)=|H|-1$. The connectivity of our dataset is turned out to be $C(G_H)\sim 0.256$, when we measured it with the equation (7). This result implies that the existence of the connectors ensures that the global social graph is not fragmented but interconnected.

V. ALGORITHMS

In this section we present two algorithms for converting a hybrid hypergraph into a canonical multigraph and a normalized weighted graph. We also introduce the problem of finding covert links and its algorithm.

A. Canonicalization and Normalization

The pseudo-codes for converting a hybrid hypergraph into canonical multigraph and normalized weighted graph are shown in Figure 4 and 5. Here we analyze the asymptotic complexity of building a global social graph, when a hybrid hypergraph is given with.

ANALYSIS 1. *Given a hybrid hypergraph* $G_H=(V, E, H)$, *the time complexity of canonicalization and normalization is* $O(|E|)$.

PROOF. It is manifest that the outer and inner loop in Algorithm 1 iterates exactly $\sum |E_i|=|E|$ times. The lookup cost for finding a hyperedge can be a constant time m, if we maintain a hashtable mapping a vertex to a hyperedge. It is also manifest that the loop in Algorithm 2 iterates exactly $|E_C|=|E|$ times, as the number of canonical edges are $|E_C|=|E|$. Hence, the total steps for running Algorithm 1 and 2 are $m|E|+c|E| = (m+c)|E|$, where c is a constant time. Therefore, the time complexity of canonicalization and normalization is $O((m+c)|E|) = O(|E|)$. □

ANALYSIS 2. *Given a hybrid hypergraph* $G_H=(V, E, H)$, *the space complexity of canonicalization and normalization is* $O(|E|)$.

PROOF. It is true that Algorithm 1 creates a new canonical edge at each step of the inner loop. Hence, the cardinality of E_C is the same as that of E, i.e., $|E_C|=|E|$. It is also true that Algorithm 2 diminishes the number of edges in E_C combining multiedges, i.e., $|E_N| \le |E_C|$, and $|E_C|+|E_N| \le 2|E|$. Therefore, the space complexity of canonicalization and normalization is $O(2|E|)=O(|E|)$. □

B. Finding Covert Links

Liben-Nowell and Kleinberg [18] formalized the link-prediction problem for social networks. Their approach was to predict future links by analyzing the proximity of nodes in a social network. In a global social graph, we can predict

ALGORITHM 3. COVERT.
INPUT: $G_C=(H, E_C)$. $e_i=(v_s, v_t) \in E_i$, $v_s, v_t \in V_i$.
OUTPUT: $S=\{ e_C \mid e_C=(v_s, v_t, G_j) \notin E_C \}$.
BEGIN.
$h_s \leftarrow$ **find a hyperedge, where** $v_s \in h_s$.
$h_t \leftarrow$ **find a hyperedge, where** $v_t \in h_t$.
$S \leftarrow \phi$.
for each $v_x \in h_s$, $v_x \neq v_s$, $v_x \in V_j$.
$v_y \leftarrow$ **find a vertex, where** $v_y \in h_t$, $v_y \in V_j$.
if (v_y **is not** *nil* **and** $e_C=(v_x, v_y, G_j) \notin E_C$) **then**
$S \leftarrow S \cup \{ e_C=(v_x, v_y, G_j) \}$
END.

Figure 6. Pseudo-codes for *covert*.

a link in a local graph using the information residing in the other local social graph. The problem of *finding covert links* can be formalized as the following: *Given a local edge,* $e_i=(v_s, v_t)$, in a site i, *find all the missing links from* h_s *to* h_t *in the other site*, *where* $v_s \in h_s$, $v_t \in h_t$.

This problem is turned out to be an initial motivation for the Google to provide the Social Graph API. To say, if a person joins a site with no friendship, covert links can be used as recommendations to provide possible links in this site, with the information residing in the other sites.

Our proposed model, a canonical multigraph makes it easy to solve this problem. Without it, in a hybrid hypergraph, finding covert links requires several operations to retrieve and match all the possible links between two users. However, in a canonical multigraph, it is made easy to find covert links from two hyperedges. Figure 6 shows the algorithm of finding all missing links when a local edge is given with. Finding a person's covert links in a new site can be done by running this algorithm for all the local edges that a person has in a hybrid hypergraph.

ANALYSIS 3. *Given a canonical multigraph* $G_C=(H, E_C)$ and a local edge $e_i=(v_s, v_t) \in E_i$, $v_s, v_t \in V_i$, *the time complexity of finding covert links is* $O(|h_s| \cdot |h_t|)$, where $v_s \in h_s$, $v_t \in h_t$.

PROOF. The loop in Algorithm 3 iterates $|h_s|$ times. The cost of finding a matching vertex in a hyperedge at each step of the loop is at most $|h_t|$, though we use a linear search. Hence, the total time complexity of finding covert links with a given local edge is $O(|h_s| \cdot |h_t|)$. □

VI. CONCLUSION

In this paper, we proposed a novel model for the global social graph as a hybrid hypergraph. We also presented two algorithms to convert it into a canonical multigraph and a weighted directed graph, and introduced three problems that make our model to be useful: finding covert links, global social routing, and global social ranking.

The contributions of this paper, based on our empirical study, are to find that (1) *the population of websites and the degree distribution of hyperedges follow the power-law*, and that (2) *the existence of connectors ensures that the global social graph is not fragmented but interconnected*.

The reason why we should study the global social graph is to understand the implications for the design of future systems based on decentralized social networks. We believe that our findings can shed a new light on these kinds of systems with a bird's-eye view of the global social graph. In our future work, we would advance to a generative model of global social graph based on power-law random bipartite graphs for simulation studies. The problems of global social routing and global social ranking should be challenged in our further studies.

REFERENCES

[1] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," In *Proc. SIGKDD'06*, Philadelphia, PA, USA, pp. 611-617, 2006.
[2] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," In *Proc. WWW'07*, Banff, Alberta, Canada, pp. 835-844, 2007.
[3] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," In *Proc. IMC'07*, San Diego, California, USA, pp. 29-42, 2007.
[4] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," In *Proc. WOSN'08*, Seattle, WA, USA, pp. 25-30, 2008
[5] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi, "Characterizing social cascades in flickr," In *Proc. WOSN'08*, Seattle, WA, USA, pp. 13-18, 2008.
[6] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," In *SIGKDD'08*, Las Vegas, Nevada, USA, pp. 462-470, 2008.
[7] H. Hu and X. Wang, "Evolution of a large online social net-work," *Physics Letters A*, vol. 373, no. 12-13, pp. 1105-1110, 2009.
[8] B. Fitzpatrick. "Thoughts on the social graph," *http://bradfitz.com /social-graph-problem*, 2007.
[9] T. Berners-Lee, "Giant Global Graph," *http://dig.csail.mit.edu /breadcrumbs/node/215*, 2007.
[10] R. Ramakrishnan and A. Tomkins, "Toward a PeopleWeb," *IEEE Computer*, vol. 40, no. 8, pp. 63-72, 2007.
[11] K. Heyman, "The move to make a social data portable," *IEEE Computer*, vol. 41, no. 4, 2008.
[12] J. Breslin and S. Decker, "The future of social networks on the internet: The need for semantics," *IEEE Internet Computing*, vol. 11, no. 6, pp. 86-90, 2007.
[13] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-442, 1998.
[14] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-512, 1999.
[15] H. Kwak, S. Han, Y. Ahn, S. Moon, and H. Jeong, "Impact of snowball sampling ratios on the network characteristics estimation: A case study of Cyworld," In *Proc. WWW'07*, Banff, Alberta, Canada, pp. 835-844, 2007.
[16] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *arXiv*: 0706.1062v2, 2007.
[17] M. E. J. Newman, "Power laws, pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, no. 5, pp. 323-351, 2005.
[18] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," In *Proc. CIKM'03*, 2003, New Orleans, LA, USA, pp. 556-559.