Eur. Phys. J. B **71**, 611–622 (2009)
DOI: 10.1140/epjb/e2009-00292-2

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Calculating statistics of complex networks through random walks with an application to the on-line social network Bebo

S.J. Hardiman[a], P. Richmond, and S. Hutzler

School of Physics, Trinity College Dublin, Dublin, Ireland

Received 18 November 2008 / Received in final form 25 May 2009
Published online 18 August 2009 –

Abstract. We develop a methodology with which to calculate typical network statistics by sampling a network through a random walk. By examining the statistics of degree and return times of walks which return to the same vertex, we can estimate characteristics of the giant component such as average clustering coefficient, degree distribution, degree correlations and network size. We confirm the validity of the methods using a variety of available network network data sets and then apply these methods to data collected by performing a random walk on the large on-line social networking website, Bebo. We find good agreement between our results and the results of previous studies of on-line social networks in which data collection was performed by a BFS ("snow-ball") sampling algorithm. In particular, we find that the degree distribution exhibits a multi-scaling power-law tail and the network exhibits clustering and positive degree correlations.

PACS. 89.65.-s Social and economic systems – 89.75.-k Complex systems

1 Introduction

Thanks to the Internet, and in particular, the recent explosion of user generated Internet content (Web 2.0), large amounts of data offering insight into the make up of complex social networks are available. They can be obtained from large user driven interactive websites such as LiveJournal [27], and in particular, popular social networking websites such as MySpace [28] and Bebo [29].

As a result, many studies into the topologies and growth of these large scale social networks have emerged in recent years. Often the researchers have had access to complete network data sets [1–4]. Other times [4,5], these studies have often been performed on data sets that have been collected from the internet using automated software that "crawls" the links between pages on these websites using a "Breadth-first search" ("snow-ball") sampling algorithm [6].

However, we aim to crawl a large on-line social network by means of a random walk and develop a methodology with which to obtain unbiased measures of social network statistics such as degree distribution, clustering coefficient and assortativity coefficient from the data collected in this way. Such a random walk crawling procedure has been used previously to assemble graphs of the World Wide Web by crawling web links [7], as well as detect communities in large networks [8,9]. We test our devised methodology on a variety of control data sets and then apply the method to random walk data collected by crawling the large on-line social network, Bebo.

[a] e-mail: hardimas@tcd.ie

2 Methodology

2.1 Degree distribution

The random walk is a time series constructed as follows: we begin by selecting a vertex v_1 randomly with uniform probability from a fully connected network. A subsequent vertex v_2 is chosen randomly with a uniform probability from the neighbours of vertex v_1. We proceed in this way for T random steps to produce a time series of vertices, $v_1, v_2, \ldots v_T$ with degrees $k_1, k_2, \ldots k_T$.

From this data set it is easy to calculate a *biased* degree distribution for the probability of visiting a vertex on the walk with degree k, $p^*(k)$,

$$p^*(k) = \frac{1}{T}\sum_{t=1}^{T} k_t \delta_{k_t,k}. \tag{1}$$

Here δ_{ij} denotes the Kronecker symbol, defined as follows

$$\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We call this distribution the biased degree distribution because the random walk does not sample vertices uniformly from the network. One can show that the probability of visiting a vertex with degree k_i on a random walk is k_i/k_j times the probability of visiting a vertex with degree k_j [10]. If we want to estimate the degree distribution for the giant component of the network, that is the fully connected bulk of the network that is accessible to

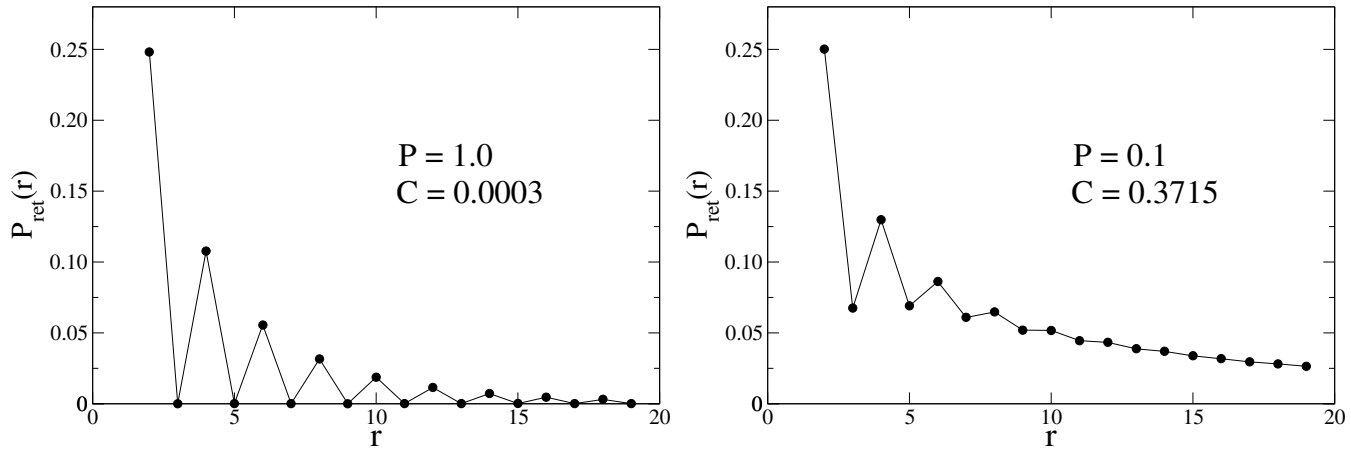

Fig. 1. $P_{ret}(r)$, the probability that we will see a return in r steps from a randomly selected vertex along a random walk through a 10 000 vertex small world network with rewiring probability P and clustering coefficient C. We calculate $P_{ret}(r)$ by counting the number of such r-returns we see in a random walk of 1 million steps. We see the shape of this function is significantly smoothed by the clustering in the network with low rewiring probability.

our random walk, then we must remove this bias, by computing the unbiased degree distribution $p(k)$,

$$p(k) = \frac{p^*(k)/k}{\sum_{l=1}^{\infty} p^*(l)/l} : k, l \in 1, 2, 3, ... \quad (3)$$

where the sum in the divisor ensures $p(k)$ is normalised to 1.

2.2 Clustering

The term "clustering" is used to describe how networks form communities of closely connected nodes. Unlike for a random network, in a social network we expect friends of friends to often also be acquainted. Although small world networks share the property of random networks of having a low mean shortest path between vertices (illustrated in the concept of "six degrees of separation"), we expect to see significant clustering due to the formation of communities, a correlation that is absent in random networks.

The clustering coefficient, C, is a network statistic which measures this clustering, and is defined as follows [11]. Let N_i be the set of vertices which connect to vertex v_i with degree k_i. At most $\frac{k_i(k_i-1)}{2}$ edges can be drawn between the elements of N_i. Let C_i be the fraction of such edges that actually exist for a given vertex i. The average clustering coefficient, C, is then defined as the average of C_i over all vertices i in the network.

We attempt to estimate the clustering coefficient of the giant component of a network by counting the number of instances in a random walk where we see a return to the same vertex in three steps. Such a 3-step return is unique because it prohibits revisiting the same nodes by taking the same route back. In a random walk through a network with a high degree of clustering, we should expect to see many more such 3-returns than, for example, 100-returns.

$P_{ret}(r)$ is the probability that, if we select a vertex randomly along the length of our walk, we will see a return to that vertex in r steps. We can calculate $P_{ret}(r)$ by counting the number of

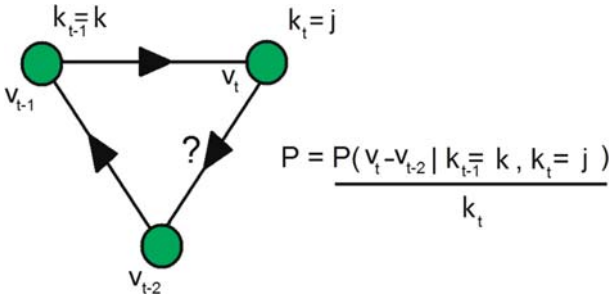

Fig. 2. (Color online) An illustration of a 3-return that may happen in the course of a random walk. The walk returns to the same vertex in 3 steps, i.e. $v_{t-2} = v_{t+1}$ with a probability given by the probability that the vertices are connected (which is a function of vertex degrees k_t and k_{t-1}) divided by the degree of vertex v_t.

observed r-returns we see during a walk of T steps and dividing by the number of possible r-returns.

$$P_{ret}(r) = \frac{\sum_{t=r+1}^{T} \delta_{v_t, v_{t-r}}}{T - r}. \quad (4)$$

In Figure 1 we have plotted $P_{ret}(r)$ for a small-world network. A small-world network is a highly clustered graph with a low mean shortest path length that we create by randomly rewiring each of the edges of an ordered network with a probability, P [11]. For $P = 0$, the network is unchanged and we retain the highly clustered ordered network. But for $P > 0$, we may introduce long-range connections which will drastically decrease the mean shortest path length of the network as well as lowering the average clustering coefficient of the network. By adjusting P, we can tune our network to a desired average clustering coefficient C.

The shape of the function, $P_{ret}(r)$, in Figure 1 can be seen to characterise the clustering behaviour of the network. For a random network (corresponding to rewiring probability $P = 1$), we see significant spikes for even values of r, corresponding to the potential for a random walk to return by the same route that it left a vertex (which is not possible in the odd-r case). In

the odd r case, $P_{ret}(r)$ takes on the constant value $P_{ret}(r) = \frac{1}{N_G}$ where N_G is the number of nodes in the giant component of the network. However, as we see in Figure 1, when we introduce clustering by decreasing the rewiring probability to $P = 0.1$, this function is significantly smoothed. Furthermore, at some arbitrarily large number, when a random walk would have lost all memory of its initial starting point, $P_{ret}(r)$ takes on the constant value $\frac{1}{N_G}$, as the probability of seeing a return in r steps becomes equal to that of seeing a return in $r+1$ steps.

Let us randomly select a vertex v_t from the random walk of T steps, $v_1, v_2, ... v_T$, such that $v_t \neq v_{t-2}$. The probability of seeing a 3-return in the next step, i.e. $v_{t+1} = v_{t-2}$, is a product of two factors, the probability that v_t and v_{t-2} are connected, and if so, the probability, $1/k_t$, that the next random step brings the walk back to vertex, v_{t-2}. It is important to note, that the probability that the vertex v_t and v_{t-2} are connected will generally be correlated with the degree of vertex v_t.

In what follows, we will write $v_a \multimap\!\circ\, v_b$ to represent that vertex v_a and v_b are connected.

Since v_t and v_{t-2} are neighbours sampled randomly and uniformly from the friends of v_{t-1}, and of all connections that can be made between neighbours of v_{t-1}, a fraction C_{t-1} exist (where C_{t-1} is the clustering coefficient of vertex v_{t-1}), the average probability that a node v_t is connected to a node v_{t-2} visited in the walk, where $k_{t-1} = k$, will be the average clustering coefficient of vertices with degree k

$$\frac{1}{N}\sum_{t=3}^{T-1}(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}c(v_t,v_{t-2}) = P(v_s \multimap\!\circ\, v_{s-2}|k_{s-1}=k) = C(k) \quad (5)$$

where $c(v_t, v_{t-2})$ specifies whether v_t is connected to v_{t-2},

$$c(v_t,v_{t-2}) = \begin{cases} 1 & \text{if} \quad v_t \multimap\!\circ\, v_{t-2} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

and N counts proper triplets of values where vertex $t-1$ has degree k,

$$N = \sum_{t=3}^{T-1}(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}. \quad (7)$$

$P(v_t \multimap\!\circ\, v_{t-2}|k_{t-1} = k)$ is a function of k, which we will call the prior probability. It is the probability that a randomly selected vertex v_t and the vertex v_{t-2}, two steps behind it, are connected, given only the knowledge that $k_{t-1} = k$. This is also the average clustering coefficient, $C(k)$, for vertices with degree k.

To calculate the expected number of 3-returns observed where $k_{t-1} = k$ during the walk, we need to know the conditional probability, $P(v_t \multimap\!\circ\, v_{t-2}|k_{t-1} = k, k_t = j)$. This is a function of both k and j which gives the probability that vertex v_t and v_{t-2} are connected given that we also know that the degree of vertex v_t is j. The expectation value for 3-returns observed is then,

$$\langle N_3\rangle_k = \sum_{t=3}^{T-1}\frac{(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}P(v_t \multimap\!\circ\, v_{t-2}|k_{t-1}=k,k_t=j)}{k_t}. \quad (8)$$

We can calculate this conditional probability using Bayes' theorem.

$$P(v_t \multimap\!\circ\, v_{t-2}|k_t=j,k_{t-1}=k) = \frac{P(k_t=j|v_t \multimap\!\circ\, v_{t-2},k_{t-1}=k)P(v_t \multimap\!\circ\, v_{t-2}|k_{t-1}=k)}{P(k_t=j|k_{t-1}=k)}. \quad (9)$$

$P(v_t \multimap\!\circ\, v_{t-2}|k_{t-1} = k) = C(k)$ is the unknown which we wish to calculate from experiment. Substituting equation (9) into equation (8) and solving for $C(k)$ gives

$$C(k) = \frac{\langle N_3\rangle_{k_{t-1}=k}}{\sum_{t=3}^{T-1}\frac{(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}P(k_t|v_t\multimap\circ v_{t-2},k_{t-1}=k)}{k_tP(k_t=j|k_{t-1}=k)}}. \quad (10)$$

What is left, is to calculate is $P(k_t = j|v_t \multimap\!\circ\, v_{t-2}, k_{t-1} = k)$ and $P(k_t = j|k_{t-1} = k)$. The latter probability is given by

$$P(k_t=j|k_{t-1}=k) = \frac{\sum_{t=3}^{T-1}(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}\delta_{k_t,j}}{\sum_{t=3}^{T-1}(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}}. \quad (11)$$

To calculate $P(k_t = j|v_t \multimap\!\circ\, v_{t-2}, k_{t-1} = k)$, we count the number of observations of vertices v_{t-1} or v_t with $k_t = j$ or $k_{t-1} = j$ from instances in the walk where $k_{t-2} = k$ and a 3-return is observed, i.e. $v_{t-2} = v_{t+1}$. This criterion samples sets of v_t, v_{t-1} and v_{t-2} for which $v_t \multimap\!\circ\, v_{t-1} \multimap\!\circ\, v_{t-2} \multimap\!\circ\, v_t$. The factor of k_tk_{t-1} and normalisation is to account for the fact that the probability of observing a 3-return through vertices with degrees k_t and k_{t-1} will be proportional to the product of these degrees.

$$P(k_t=j|v_t \multimap\!\circ\, v_{t-2},k_{t-1}=k) = \frac{\sum_{t=3}^{T-1}\delta_{v_{t+1},v_{t-2}}\delta_{k_{t-1},j}\delta_{k_t,j}k_tk_{t-1}}{\sum_{t=3}^{T-1}\delta_{v_{t+1},v_{t-2}}k_tk_{t-1}}. \quad (12)$$

We can now estimate $C(k)$ using equation (10) and our experimental value of $N_3(k)$ found by counting the number of 3 returns observed during the random walk.

$$C(k) = \frac{N_3(k)}{\sum_{t=3}^{T-1}\frac{(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}P(k_t=j|v_t\multimap\circ v_{t-2},k_{t-1}=k)}{k_tP(k_t=j|k_{t-1}=k)}}$$
$$N_3(k) = \sum_{t=3}^{T-1}(1-\delta_{v_t,v_{t-2}})\delta_{k,k_{t-1}}\delta_{v_{t+1},v_{t-2}}. \quad (13)$$

By averaging $C(k)$ over the unbiased probability distribution $p(k)$ of equation (3), we can estimate the average clustering coefficient for the giant component of the network, C.

$$C = \sum_{k=1}^{\infty}p(k)C(k). \quad (14)$$

2.3 Degree correlations

It has been argued that social networks, distinct from technological and biological networks show positive degree correlations, also called assortative mixing between the degrees of adjacent vertices [12]. People with many friends tend to have friends with many friends. Or in different words, popular people associate with other popular people.

One way to inspect the degree correlations is to examine the average degree, $\langle k_{nn}\rangle$, of the neighbours of a vertex as a function of vertex degree k. For an assortative network, we should expect $\langle k_{nn}\rangle(k)$ to be an increasing function of k. Another measure of assortativity can be obtained with the assortativity coefficient [13],

$$r = \frac{1}{\sigma^2}\sum_{k_1,k_2} k_1 k_2 (p_e(k_1,k_2) - p_e(k_1)p_e(k_2)) \tag{15}$$

where $p_e(k)$ is the probability of finding a vertex with degree k as one of the ends of an edge chosen randomly from the network. $p_e(k_1,k_2)$ is then the joint probability for finding vertices of both degree k_1 and k_2 at the ends of a randomly chosen edge. r is normalised by the variance, σ^2 of $p_e(k)$, given by

$$\sigma^2 = \sum_k k^2 p_e(k) - \left(\sum_k k p_e(k)\right)^2. \tag{16}$$

We approximate sampling edges uniformly from a network using the sample space of edges visited on a random walk, $p_e(k) \approx p^*(k)$. The assortativity coefficient then simply becomes the 1-step auto-correlation function of the time series of degrees visited on the random walk. That is,

$$r = \frac{1}{\sigma^2}\frac{1}{T}\sum_{t=1}^{T-1}\langle k_t k_{t+1}\rangle - \langle k_t\rangle\langle k_{t+1}\rangle \tag{17}$$

where

$$\langle k_t\rangle = \frac{1}{T}\sum_{t=1}^{T} k_t \qquad \sigma^2 = \frac{1}{T}\sum_{t=1}^{T}\langle {k_t}^2\rangle - \langle k_t\rangle^2. \tag{18}$$

Also, using the edges that we see during the random walk, we can calculate the average neighbour degree of vertices as a function of vertex degree.

$$\langle k_{nn}(k)\rangle = \frac{\sum_{t=1}^{T-1} k_{t+1}\delta_{k,k_t} + \sum_{t=2}^{T} k_{t-1}\delta_{k,k_t}}{\sum_{t=1}^{T-1}\delta_{k,k_t} + \sum_{t=2}^{T}\delta_{k,k_t}}. \tag{19}$$

2.4 Size of giant component

The total number of vertices, N_G, in the giant component of a network can be inferred by counting the number of visits to a given vertex that we see during a random walk. In what follows, we use $\langle\rangle$ to represent a uniform average over vertices visited in the giant component of the network, and $\langle\rangle^*$ to represent an average over vertices visited in a random walk, which as already stated, is biased to visit vertices with large degree.

In Figure 3, we show the cumulative sum, $F(r)$, of $P_{ret}(r)$ for a network, defined as,

$$F(r) = \sum_{t=1}^{r} P_{ret}(t). \tag{20}$$

Beyond a certain critical value, r_c, we see linear behaviour in the plot of $F(r)$. This corresponds to $P_{ret}(r) \approx P_{ret}(r+1)$ for sufficiently large r, indicating that it is just as likely to return to a vertex after 100 steps, as it is after 101 etc. The probability, $P^\infty(k)$ that an arbitrarily large number of steps finds us at a given vertex with degree k, is given by [10]

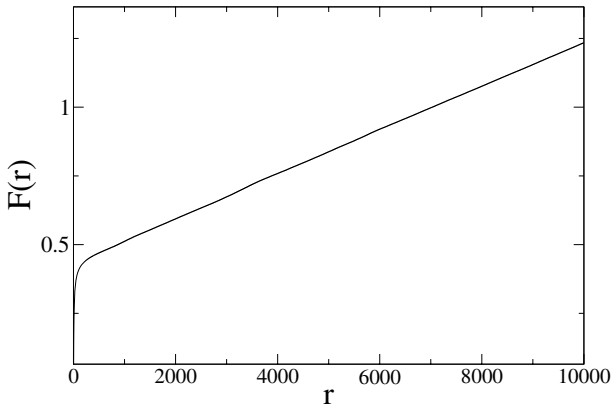

Fig. 3. Cumulative sum of $P_{ret}(r)$ for the condensed matter collaboration network [14] as calculated by equation (20). We see linear behaviour for $r > 300$ corresponding to $P_{ret}(r) = P_{ret}(r+1)$ for large r. This shows that it is just as likely to return to a vertex after r steps, as it is after $r+1$ and allows us to compute the value of $\langle P^\infty\rangle^* = 0.0000801$ as the slope of $F(r)$.

$$P^\infty(k) = \frac{k}{\sum_i^{N_G} k_i} = \frac{k}{N_G\langle k\rangle}. \tag{21}$$

Let us call the average of $P^\infty(k)$ over all vertices visited in our walk $\langle P^\infty\rangle^*$,

$$\langle P^\infty\rangle^* = \sum_{k=1}^{\infty} p^*(k)P^\infty(k). \tag{22}$$

Inserting equation (21) into equation (22), we obtain

$$\langle P^\infty\rangle^* = \frac{\langle k\rangle^*}{N_G\langle k\rangle} \tag{23}$$

$$\Rightarrow N_G = \frac{\langle k\rangle^*}{\langle P^\infty\rangle^*\langle k\rangle}. \tag{24}$$

Thus, given $\langle P^\infty\rangle^*$, the average degree of vertices visited in the walk, $\langle k\rangle^*$, and the average degree of vertices in the giant component, $\langle k\rangle = \sum_{k=1}^{\infty} p(k)k$, we can estimate the number of vertices, N_G, in the giant component of the network. $\langle P^\infty\rangle^*$ can be taken from the slope of $F(r)$ for $r > r_c$.

3 Results on network data sets

3.1 Degree distribution

In Figure 4 we use equation (3) to estimate the degree distribution of a selection of networks using data collected in random walks. The walk lengths used represent only a small fraction of the number of vertices in the giant component of the networks, yet we find very good agreement between our estimates and the degree distributions as calculated directly from the data.

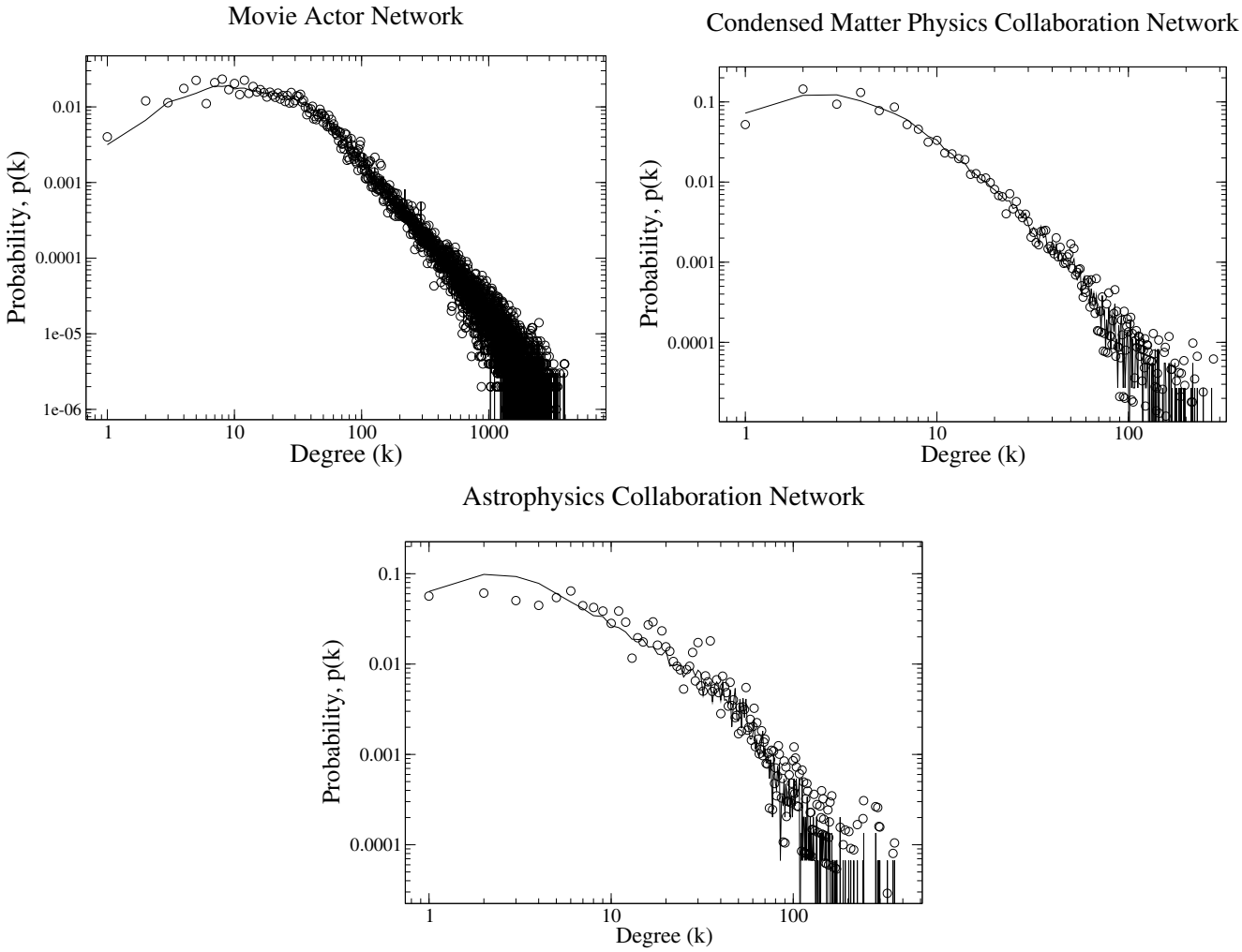

Fig. 4. Estimate (circles) of giant component degree distributions for three network data sets using equation (3) compared against the degree distribution as calculated directly from the data (line). (a) movie actor network [15]: $N_G = 374\,550$, walk length = 40 000, vertices sampled = 31 100. (b) condensed matter collaboration [14]: $N_G = 14\,845$, walk length = 5000, vertices sampled = 3228. (c) astrophysics collaboration [14]: $N_G = 36\,458$, walk length = 2000, vertices sampled = 1432.

3.2 Clustering

In Figure 5 we have plotted the estimate of the average clustering coefficient as a function of degree k as estimated using the random walk procedure for a network data set. Since 3-returns are rare, we require very long random walks to yield enough 3-returns to provide sufficient accuracy. Since our result for the clustering coefficient is calculated from the count of 3-returns observed in our random walk. we expect the error in our result to be at least as large as the random counting error, that is

$$\frac{\Delta C}{C} \geq \frac{\sqrt{N_3}}{N_3}. \tag{25}$$

In Figure 5 we use equation (13) to estimate the average clustering coefficient as a function of vertex degree for the condensed matter physics collaboration Network [14] using a 1 million step long random walk, and compare it to the result as calculated directly from the whole data set. To ensure a large number of three returns for demonstrating the validity of equation (13), a walk length, T, was chosen such that $T \gg N_G = 36,458$, the size of the giant component of the network.

In Table 1, we have used equation (14) to estimate the clustering coefficient for a selection of network data sets, using very long walk lengths and compared the results to exact values for the giant components as calculated directly from the data. We find very good agreement between the exact values, and the estimates.

To investigate how the error in the result depends on walk length, in Figure 6, we have estimated the clustering coefficient of the giant component of the condensed matter network as a function of the length of random walk for 100 hundred unique walks and plotted the standard deviation of their average. We find the standard deviation of the estimate to decrease roughly as one over the square root of the walk length or as one over the square root of the number of 3-returns observed, much like random counting error.

It is clear that if we want to estimate the clustering coefficient accurately by counting the 3-returns in this way, we require the walk to run long enough to count enough 3-returns

Table 1. Clustering coefficient estimates from random walks and exact values.

Network	Size of giant component	Walk length	3-returns counted	Vertices visited	Avg. C of giant component	Estimated C of giant component
Condensed matter [14] Collaboration	36 458	1 000 000	34 677	35 939	0.656585	0.671141
High energy physics [14] Collaboration	5835	1 000 000	52 323	5835	0.506193	0.505856
Astrophysics [14] Collaboration	14 845	1 000 000	23 518	14 771	0.669618	0.657160
Protein interaction [16] Network	1458	100 000	1185	1458	0.070830	0.068152
Email correspondence [17] Network	1133	100 000	1241	1133	0.220176	0.217081
Small world [11] Network	30 000	1 000 000	3734	30 000	0.146340	0.145449
Actor collaboration [15] Network	374 550	20 000 000	105 990	370 402	0.778739	0.765048

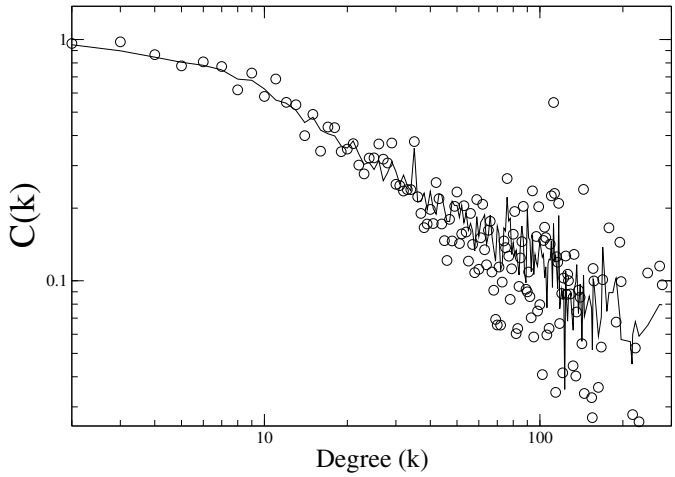

Fig. 5. Average clustering coefficient (circles) as a function of vertex degree as estimated using equation (10) from a one million step long random walk through the condensed matter physics collaboration network [14]. This result is compared with the exact value for $C(k)$ (line) as calculated using the exact network data.

with which to make a reliable average. This method is therefore going to be best suited to networks with a small ratio of number of edges to number of nodes. Such networks will have a low average degree, making 3-returns more likely to occur. The method may also be a suitable choice for calculating an unbiased estimate of C for a network for which the full data set is unavailable, but for which the network may be too large to realistically cover with a "snow-ball" sampling approach.

3.3 Degree correlations

In Figure 7 we show the random walk estimate of $\langle k_{nn} \rangle(k)$ for the giant component of the condensed matter physics collaboration network using equation (19) with the exact function as calculated directly from the data. We find very good agreement. This shows that using the sample space of edges visited on the random walk is a good approximation to sampling edges uniformly from the network.

In Table 2 we tabulate our estimates for the assortativity coefficient for a variety of networks and walk lengths, and compare these to the exact assortativity coefficient as calculated for the full network data sets. We find very good agreement, even in the cases where the walk lengths are short when compared to the full size of the giant components.

3.4 Size of giant component

In Table 3 we have performed random walks on a selection of networks and used equation (24) to estimate the sizes of the giant components. We find the method gives very reliable estimates, even in the cases when the random walks are small compared to network size.

4 The Bebo network

In this section, we apply the methods described and tested in the previous two sections to a random walk performed on a large on-line social network, Bebo, for which the full network data set is not available to us.

A social networking website, or an SNS (Social Networking Service) is a relatively recent phenomena. It is an interactive website that allows visitors to set up their own personal web-page which serves as a base from which they can communicate and share media with other visitors to the website. An important aspect of a social networking website is the ability to add "friends" to one's profile web-page. This process creates a direct *HTML* hyperlink from the member's profile web-page to the profile web-page of said friend. These friends are

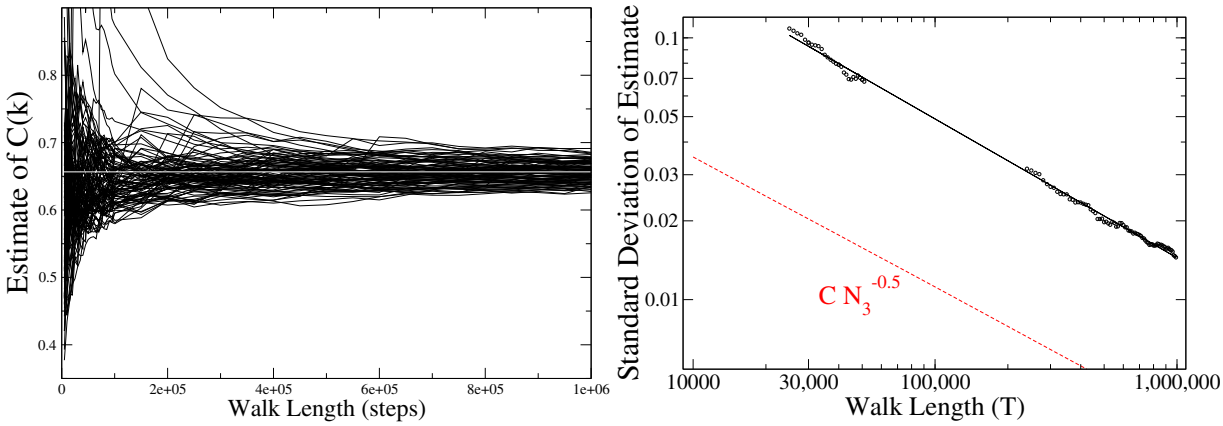

Fig. 6. (Color online) Left: estimate of $C(k)$ for giant component of the condensed matter collaboration network as a function of random walk length, for 100 different realisations of a random walk. Right: standard deviation of these estimates as a function of walk length and the lower bound on this expected error by analogy with random counting error.

Table 2. Assortativity coefficient estimates from random walks and exact values.

Network	Size of giant component	Walk length	Assortativity coefficient of giant component	Estimated assortativity coefficient of giant component
Condensed matter [14] Collaboration	36 458	1 000 000	0.177385	0.176426
High energy physics [14] Collaboration	5835	8000	0.185162	0.178474
Astrophysics [14] Collaboration	14 845	2000	0.227751	0.200222
Protein interaction [16] Network	1458	500	–0.209541	–0.222170

Table 3. Giant component size estimates from random walks and exact sizes.

Network	Size of giant component	Walk length equation (20)	Slope of $F(r)$	Avg. degree of walk $\langle k\rangle^*$	Estimated Avg. degree of giant component $\langle k\rangle$	Estimated size of giant component
Condensed matter [14] Collaboration	36 458	1 000 000	8.01×10^{-5}	27.65	9.33	36 998
Astrophysics [14] Collaboration	14 845	5000	2.0×10^{-4}	46.40	16.83	13 784
Email [17] Correspondence	1133	200	1.75×10^{-3}	18.55	9.54	1111
Movie actor [15] Collaboration	374 550	10 000	1.55×10^{-5}	431.28	82.68	336 532

most often real world acquaintances (school friends, college friends, workmates etc.) who have also subscribed to the website and established their own member page. As a result, networks made through ties on on-line social networking websites provide a good comparison to real world networks of acquaintances. Friendship on Bebo is a symmetric relation (if A is a friend of B, then B is a friend of A) and since no information exists for the strength of the personal ties, the network is undirected and unweighted.

Bebo, however, is a very large scale network, and due to the time-consuming nature of the crawling process, we perform our data collection by probing only a small fraction (~1%) of its accessible pages. However, with the postulate that the Bebo network is homogeneous and well-connected (every vertex can be reached from almost every other vertex by a short path), we are confident that a freely wandering random walk sampling procedure can produce a more uniform and representative data set, than a similar sized, but more localised "snowball" sample

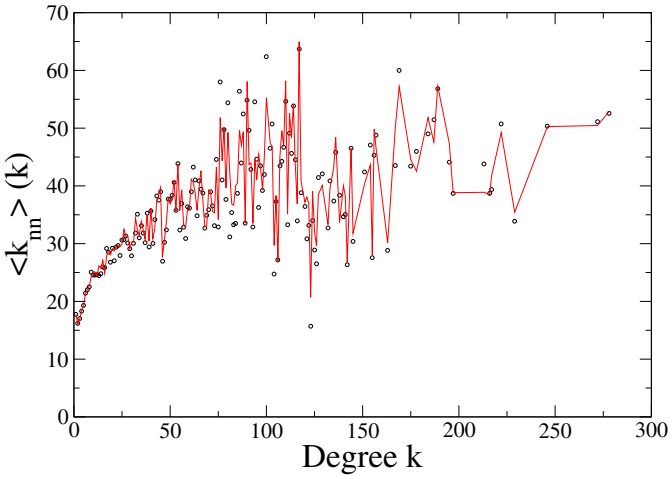

Fig. 7. (Color online) Average degree of neighbouring vertices as a function of vertex degree for the giant component of the condensed matter collaboration network. Estimate calculated from 100 000 step random walk shown in circles, exact value as calculated from the full data set shown by the line.

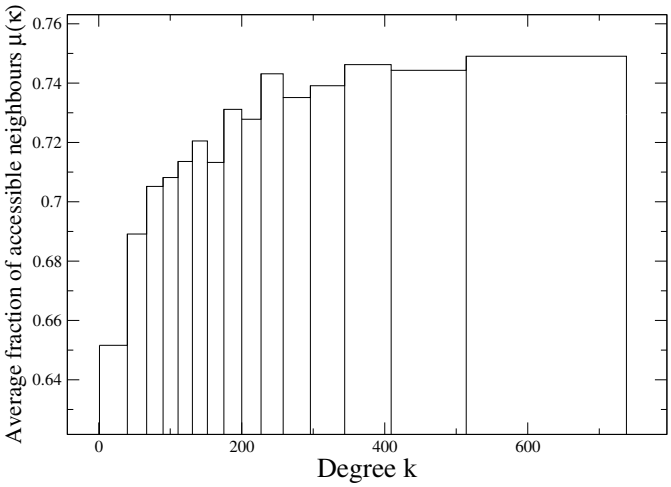

Fig. 8. Accessibility of neighbours in the Bebo network. By counting the number of times we are granted or denied access when taking a random hop from a vertex of degree k, we estimate the form of the function $\mu(k)$, the average fraction of a vertex's neighbours that have accessible profiles. The width of the bins were chosen such that approximately the same number of points (hits/misses), 13 500, were used in the calculation of the $\mu(k)$ for each k range.

of the network. Unfortunately, statistics such as average path length between vertices and vertex centrality cannot be sampled using this random walk procedure.

4.1 Data collection

The data collection was performed as follows: A random member page was selected, its unique member number identified and its degree (number of Bebo friends as stated on the profile page) recorded. We then visited the Bebo page of a randomly selected friend from the initial member page, again recorded member number and degree and made another random choice amongst its stated friends. This random walk through the Bebo network was continued for over 150 000 steps. We automated this process by developing software similar to that of web crawlers used by search engines to collect data about the World Wide Web. It resulted in the recording of a large data set of unique member ID and degree. The crawl was performed over 10 days between April 15th 2008 and April 24th 2008, during which period we believe the 3 year old network not to have changed significantly.

Since some Bebo members, for privacy reasons, choose not to have their profile visible to the public, our random walk can only peruse a sub-network of the Bebo network composed of *accessible* profile pages. Moreover, it is not possible to tell from the information available on a given Bebo member profile page, how many of that user's friends have private profiles, without performing an exhaustive search that attempts to visit the profile page of each and every friend to check if access is granted or denied. Therefore a random walk will often hit upon member pages which are inaccessible. When this occurs, we randomly select a different friend and continue. We count the number of "hits" and "misses" that we see on our walk, after leaving a vertex with total degree k, to form an estimate for the function $\mu(k)$, the average fraction of friends of a vertex with total degree k who have publicly accessible profiles. The form of this function is shown in Figure 8. We can only *estimate* the degree of a given vertex with respect to this accessible sub-network using the function $\mu(k)$. Therefore a distinction must be made between the degree of a vertex as stated on the Bebo user's profile page, k, and the estimated accessible degree of that vertex $\mu(k)k$. Thus, in the application of the described methodology to the Bebo network, certain equations, such as equations (3) and (8), need to be modified to use instead of k, the estimate of the accessible degree, $\mu(k)k$. This is an unaddressed source of error in the analysis below.

4.2 Degree distribution

In Figure 9 we show the normalised biased and unbiased degree distributions, $p^*(k)$ and $p(k)$ as calculated using equations (1) and (3) for the Bebo network.

Complex networks such as the world wide web [18,19], biological [16,20] and ecological [21] networks have been shown to exhibit scale-free nature characterised by a power law degree distribution. Such power law degree scaling has been attributed to models of growth and preferential attachment, as exemplified by the Barabsi-Albert model [22,23]. Evidence also exists to support that many social networks show power law scaling behaviour such as in the movie actor collaboration network [22], in networks of scientific collaboration [23] and in mobile communications networks [24].

To investigate power law scaling in the Bebo social network, we have plotted, in Figure 10 the cumulative degree distribution $p_{>k}(k)$, given by

$$p_{>k}(k) = \sum_{l=k}^{\infty} p(l). \tag{26}$$

This function, $p_{>k}(k)$ gives the probability of having more than k friends. A power law decay, $p(k) \sim k^{-\gamma}$, in the tail of the

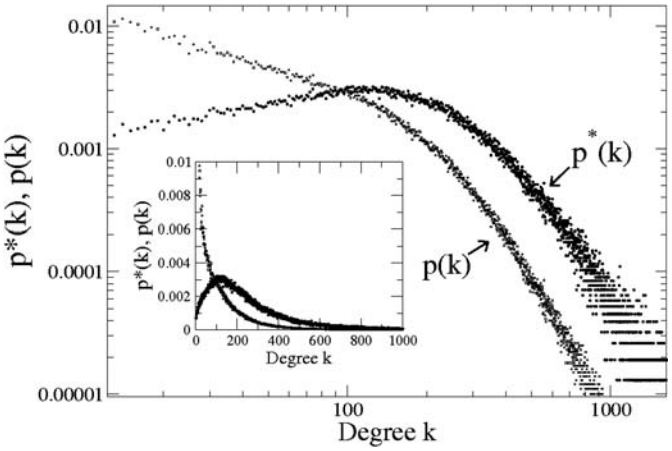

Fig. 9. $p^*(k)$ is the biased degree distribution for the Bebo network calculated by counting the degrees of vertices visited on the random walk. $p^*(k)$ is biased towards towards higher k since the probability of visiting a vertex on a random walk through a network is proportional to its degree. $p(k)$ is the unbiased estimate of the degree distribution for the giant component of the Bebo network calculated using equation (3). Functions shown on a log-log scale with the inset on a linear scale.

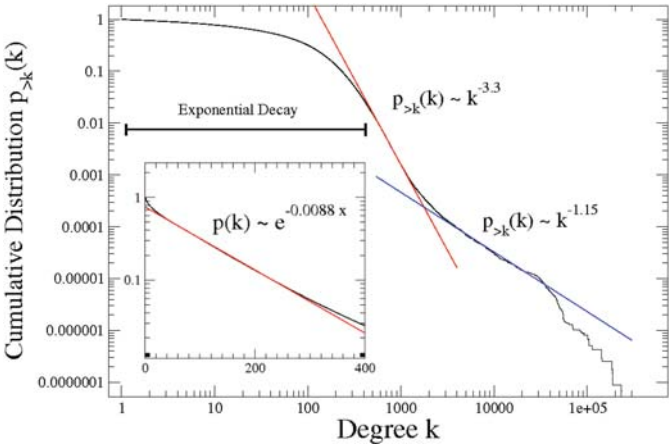

Fig. 10. (Color online) The cumulative degree distribution, $p_{>k}(k)$ as defined in equation (26). We see two power law regimes in the tail described by fits with $p_{>k}(k) \sim k^{-3.3}$ and $p_{>k}(k) \sim k^{-1.15}$. The latter regime is most likely due to "collectors" and not representative of real social ties. The inset shows the region of this distribution between $k = 0$ and $k = 400$ on a log linear scale where the function is well described by an exponential decay.

degree distribution $p(k)$, with exponent γ will be exhibited by a power law decay in the cumulative, $p_{>k} \sim k^{-\gamma+1}$.

We find that neither a log-normal distribution or single power law distribution are adequate to describe our data, and characterise the degree distribution as such: for $k < 500$, the cumulative degree distribution, shown in Figure 11, is described by an exponential decay with decay constant −0.0088. For values of k in the region $500 \leq k < 2500$, the exponential decay gives way to a power law decay, $p_{>k}(k) \sim k^{-\gamma}$ with exponent $\gamma = 3.3$. For $2500 \leq k \leq 30\,000$ we see a second power-law region with an exponent $\gamma = 1.15$.

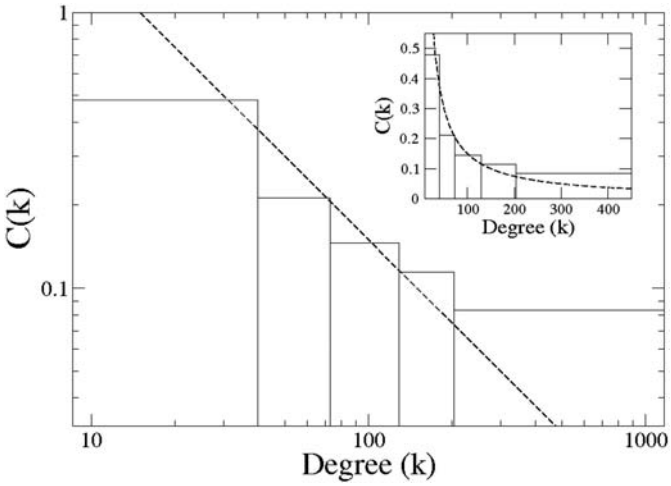

Fig. 11. Estimate of the average clustering coefficient, $C(k)$, as a function of vertex degree for the Bebo network on a log-log scale (with the inset on a linear scale.) The dashed line varies as $\sim k^{-1}$. Since only 193 3-returns were recorded during the random walk, this graph is binned to a limited resolution. The bin sizes were chosen to reflect the distribution of 3-returns ($v_t = v_{t-3}$) with respect to the degree of vertices v_{t-2} such that $193/5 \approx 39$ 3-returns were used in the calculation of each bin. As such we can expect a fractional error of at least $39^{-0.5}$ for each bin.

Since profile pages with degree $k > 2500$ are probably, in most cases, too large to represent real personal relationships, we suggest that the second power law exponent is representative of a different type of Bebo user who's goal is to "collect" on-line friends. These friends are not representative of real social ties, as we would expect for small k. Finally, in the region where $k > 30\,000$ the data is sparse and as such, unreliable. Moreover, in most cases, such data is not representative of actual Bebo users, but gimmicks to which Bebo members subscribe by adding as a friend.

Such a multi-scaling behaviour with two power-law regions has been observed before in the degree distribution of the Cyworld network, a Korean SNS website similar to Bebo [4]. In the cumulative degree distribution for the full Cyworld SNS network there are rapidly decaying and heavy tailed power-law regions with power law exponents $\gamma \sim -4$ and $\gamma \sim -1$, similar to the value of $\gamma \sim -3.3$ and $\gamma \sim -1.15$ that we find for the Bebo network. This suggests that a multi-scaling degree distribution may be a universal feature of the complex networks of large on-line SNS websites.

4.3 Clustering coefficient

Since the average degree of the Bebo network is quite high, during our walk of approximately 150 000 random steps we only observe 193 3-returns. We have therefore binned our estimate of $C(k)$ to a very limited resolution, this is shown in Figure 11. In many complex networks the clustering coefficient as a function of degree is found to behave as $C(k) \sim k^{-1}$, indicating a hierarchical topology [25]. To investigate this phenomenon in the case of the Bebo network, we have plotted a dashed line in Figure 11 which varies as $\sim k^{-1}$.

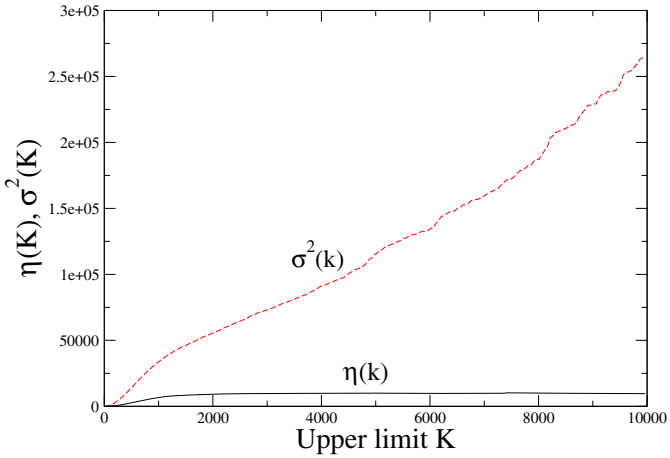

Fig. 12. (Color online) The numerator (Eq. (27), solid line) and divisor (Eq. (28), broken line) of the assortativity constant are plotted separately, and as a function of the limiting degree in their respective summations. The numerator (covariance) settles after $K = 1000$. However due to the power law tail of the degree distribution function, the denominator (variance) diverges with increasing K, demonstrating that the assortativity coefficient is ill-defined on a scale-free network.

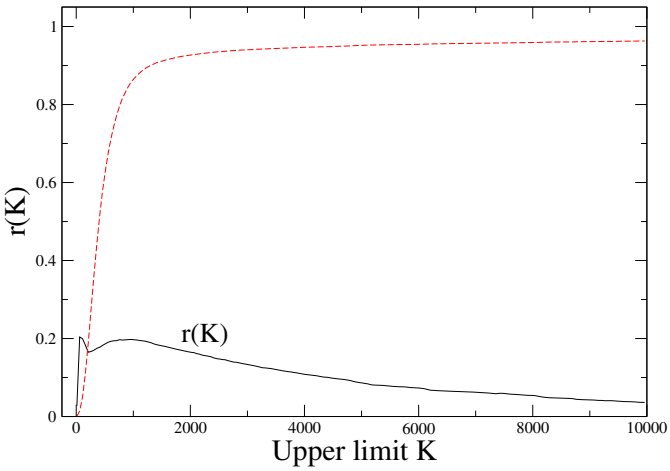

Fig. 13. (Color online) The assortativity constant, r (solid), as a function of the limiting degree in the sum, K, equation (29). $r(k)$ is roughly constant for $100 < k \leq 1000$ and then decays for $k > 1000$ due to the diverging variance shown in Figure 12. The dashed line shows the fraction of the walk composed of connected vertices each with degree $k < 1000$.

When we average this function $C(k)$ over the unbiased degree distribution $p(k)$ for the Bebo network, we get an estimate for the average clustering coefficient for the giant component of the Bebo network to be 0.248. With only 193 3-returns and a poor approximation to the function $C(k)$, the error in this estimate is likely to be large. However, it is worth noting that when we shuffle the data set and break the correlations present in the random walk time series, on average we only count 0.25 3-returns and estimate C to be 0.0001.

4.4 Degree correlations

Naively applying equation (17) to the data collected on the Bebo random walk, gives a value for the assortativity coefficient of –0.004243. It should be noted that a difficulty arises when applying this measure of assortativity to a network with a power law tail since $p(k)$ should have infinite variance. $r(k)$ is therefore an ill-defined measure on such a network. We can of course estimate a variance for the Bebo network from our finite sample, but due to the power law tail of $p^*(k)$, the σ^2 that we calculate in equation (16) is heavily dependent on the limit of the sum over k. We will get different values of r depending on the value of this limit. This is illustrated in Figures 12 and 13, where

$$\sigma^2(K) = \sum_{k=1}^{K} k^2 p_K^*(k) - \left(\sum_{k=1}^{K} k p_K^*(k) \right)^2 \quad (27)$$

$$\eta(K) = \sum_{k_1,k_2}^{K} k_1 k_2 (p_K^*(k_1,k_2) - p_K^*(k_1) p^*(k_2))) \quad (28)$$

$$r(k) = \frac{\eta(K)}{\sigma^2(K)}. \quad (29)$$

$p_K^*(k)$ and $p_K^*(k_1,k_2)$ are the normalised degree distribution, and joint degree distribution when restricting the sample space to only connected pairs of vertices each of which have degree $k < K$.

In Figure 13, we see that $r(k)$ is roughly constant for $100 < K < 1000$, and then slowly decays for increasing K due to the influence of the steadily diverging variance, $\sigma^2(K)$, shown in Figure 12. For this reason we believe the best estimate for $r(k)$, the assortative coefficient of the Bebo network, is calculated by ignoring neighbours with degree $K > 1000$ in the sum of equation (15). We believe 1000 is a reasonable cut off point, since pairs of vertices each with degrees <1000 represent 86.4% of connected pairs from the walk as illustrated in Figure 13. This gives us a value for the assortative coefficient of 0.21. If we perform the same analysis on data sets formed by shuffling the order of our random walk, we find an average assortativity constant of 0.0 with a standard deviation of 0.007. In Figure 14 we use equation (19) to show the average neighbour degree as a function of vertex degree for the Bebo network. Here an upward trend again confirms the assortative nature of the Bebo network.

4.5 Size of giant component

In Figure 15. We show the function $F(r)$ for the Bebo network as calculated using equation (20). We estimate the slope to be $\langle P^\infty \rangle^* = 1.21 \times 10^{-6}$. With the average accessible degree of vertices visited on the walk, $\langle \mu(k)k \rangle^* = 67.7$, and the estimated average accessible degree for the giant component, $\langle \mu(k)k \rangle = 575.3$, we can estimate the number N_G of vertices in the accessible sub-network,

$$\Rightarrow N_G \approx \frac{575.3}{67.7 \langle P^\infty \rangle^*} \approx 7.07 \times 10^6.$$

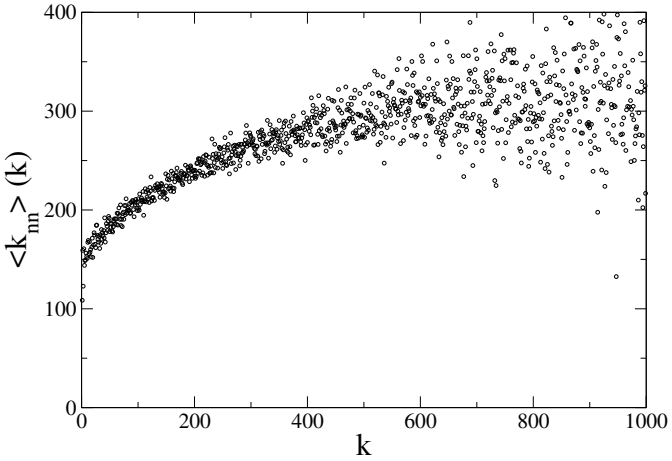

Fig. 14. Average neighbour degree $\langle k_{nn} \rangle$ as a function of vertex degree. Neighbours with degree $k > 1000$ are not included in the average. The increase of $\langle k_{nn} \rangle$ with k confirms the assortative nature of the accessible Bebo sub-network.

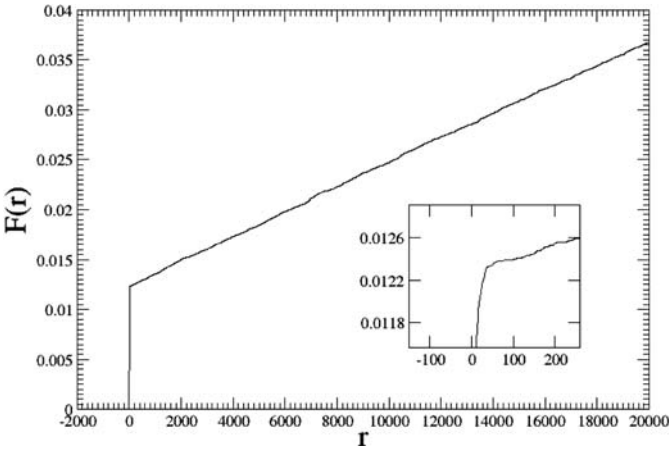

Fig. 15. Cumulative sum of $P_{ret}(r)$ as calculated by equation (20). We see linear behaviour for $r > 100$ corresponding to $P_{ret}(r) = P_{ret}(r+1)$ for large r. This shows that it is just as likely to return to a vertex after r steps, as it is after $r+1$ and allows us to compute the value of $\langle P^{\infty} \rangle^{*}$ as the slope of $F(r)$.

5 Summary and conclusions

We have proposed a method for the computation of common statistical measures of complex networks by means of performing random walks on the networks. To test the methodology we used a variety of control data sets, and find good agreement with exact statistics as calculated by the normal means. We then apply the method to a large on-line social network, Bebo.

From the data collected on a random walk through the publicly accessible profile web pages of the web-based social network Bebo, we have estimated the degree distribution of its giant component. We find that the degree distribution exhibits a multi-scaling behaviour that can be characterised by two power-law exponents. This is perhaps indicative of two types of Bebo members, those who's friends are representative of real social ties, and those who use the Bebo network simply to "collect" friends.

We count a significant number of "3-returns" in the data, strong evidence for clustering, confirming that members of the website tend to form communities of closely connected friends. However, with only 193 returns recorded, we cannot produce a function for $C(k)$ with sufficient accuracy to reliably estimate a clustering coefficient for the giant component of the network.

We also see positive degree correlations, resulting in a positive assortativity constant of 0.21, and thus confirming the notion that popular people befriend other popular people on Bebo. These results are in line with previous studies on social networks [12].

By looking at the frequency of returns to vertices on the random walk, we *estimate* the size of the Bebo network to be 7.07×10^6. Although this figure for the size is of the correct order of magnitude, it disagrees with the value published in the press and popular media of about 40×10^6 [26].

It is important to stress, that the random walk is only able to explore the publicly accessible portion of the entire Bebo network, and our figure of 7×10^6 is only representative of a fully connected giant component of publicly accessible vertices. It is likely that a significant number of the users of Bebo may have inaccessible profiles or exist on small sub-networks disconnected from the giant component. This sampling procedure also assumes that the giant component of the Bebo network is well connected, but it could be possible that many profiles or islands of profiles can only be reached by long chains of connections, which are difficult to sample by means of a random walk.

The walk is also by no means extensive; the profile pages visited represent only about 1% of the total Bebo network. Nevertheless we believe that our data characterizes this network sufficiently. Furthermore our developed methodology should be of benefit to future related studies.

This work is supported the Irish Research Council for Science, Engineering and Technology: funded by the National Development Plan and the European Science Foundation Cost actions P10, "Physics of Risk" and MP0801 "Physics of Competition and Conflicts".

References

1. R. Kumar, J. Novak, A. Tomkins, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (2006), pp. 611–617
2. J. Leskovec, L. Backstrom, R. Kumar, A. Tomkins, Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (2008), pp. 462–470
3. F. Fu, L. Liu, L. Wang, Physica A **387**, 675 (2008)
4. Y.Y. Ahn, S. Han, H. Kwak, S. Moon, H. Jeong, Proceedings of the 16th international conference on World Wide Web (2007), pp. 835–844
5. A. Mislove, H.S. Koppula, K.P. Gummadi, P. Druschel, B. Bhattacharjee, Proceedings of the first workshop on Online social networks (2008), pp. 25–30
6. S.H. Lee, P.J. Kim, H. Jeong, Phys. Rev. E **73**, 016102 (2006)

7. M.R. Henzinger, A. Heydon, M. Mitzenmacher, M. Najork, Proceedings of the 9th international Conference on World Wide Web (2000), pp. 295–308
8. P. Pons, M. Latapy, Lect. Notes Comput. Sci. **3733/2005**, 284 (2005)
9. R. Lambiotte, J.C. Delvenne, M. Barahona (2009), e-print arXiv:0812.1770
10. J.D. Noh, H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004)
11. D.J. Watts, S.H. Strogatz, Nature (London) **393**, 440 (1998)
12. M.E.J. Newman, J. Park, Phys. Rev. E **68**, 036122 (2003)
13. G. Caldarelli, *Scale-Free Networks, complex webs in nature and technology*, Oxford University Press (2007), p. 27
14. M.E.J. Newman, Proc. Natl. Acad. Scie. USA **98**, 404 (2001)
15. A.L. Barabási, R. Albert, Science **286**, 509 (1999)
16. H. Jeong, S.P. Mason, A.L. Barabási, Z.N. Oltvai, Nature **411**, 41 (2001)
17. R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, Phys. Rev. E **68**, 065103 (2003)
18. R. Albert, H. Jeong, A.L. Barabási, Nature (London) **401**, 130 (1999)
19. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Computer Networks **33**, 309 (2000)
20. H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabási, Nature **407**, 651 (2000)
21. G. Caneva, M. Cutini, A. Pacini, M. Vinci, Plant Biosystems **136**, 291 (2002)
22. A.L. Barabási, R. Albert, Science **286**, 509 (1999)
23. A.L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, Physica A **311**, 590 (2002)
24. J.P. Onnela, J. Saramaki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.L. Barabási, Proc. Natl. Acad. Sci. **104**, 7332 (2007)
25. E. Ravasz, A.L. Barabási, Phys. Rev. E **67**, 026112 (2003)
26. International Herald Tribune, AOL to buy social networking site Bebo (March 13, 2008)
27. http://www.livejournal.com/
28. http://www.myspace.com/
29. http://www.bebo.com/