Structural Analysis of Social Networks with Wireless Users

Guanling Chen and David Kotz

{glchen, dfk}@cs.dartmouth.edu

Dartmouth College Computer Science Technical Report TR2005-549

July, 2005

Abstract

Online interactions between computer users form Internet-based social networks. In this paper we present a structural analysis of two such networks with wireless users. In one network the wireless users participate in a global file-sharing system, and in the other they interact with each other through a local music-streaming application.

1 Introduction

Human interactions can be modeled as *social networks*, in which a node represents a person and an edge indicates some interaction between two people. An interaction between two persons may take many forms: they are relatives, they went to school together, they have co-authored some paper, and so on. As the Internet becomes ubiquitous, people's online interactions have also become an important part of modern social life. A good example is the email network on which people increasingly rely for their communication needs.

One of the goals among social network analysts is to find ways of revealing the underlying network structures. Researchers have studied different social networks and found some interesting patterns [4, 15, 5, 6]. These results have strong implications on how to build robust and fault-resilient systems [3], how to design more efficient search algorithms [1, 9], and how to contain epidemic viral propagations [11]. For instance, protecting the highly-connected nodes on the Internet is critical to minimize the impact of epidemic outbreak.

Recent advances in Wireless LAN (WLAN) technologies enable end users to use many applications on increasingly-powerful mobile devices. The interaction patterns of wireless users are likely to differ than those using stationary desktops, due to user mobility, battery constraints, and device form factors. Building software tools that create, maintain, and optimize social networks in such a wireless environment requires understanding of network characteristics and how the network evolves.

In this paper we present a statistical analysis of the structures of two online social networks, formed by wireless users' interactions through peer-to-peer applications of global file sharing and local music streaming. By analyzing a four-week WLAN trace, we found that the wireless users in the global file-sharing network had few local communications. The majority of both networks were well-connected, though the global file-sharing network had a heavy-tailed distribution of the wireless node degree and the local music-streaming network did not. The local music-sharing network had both low clustering and low average shortest path length between any two nodes. We discuss the implications of these results on content sharing, distributed query, and worm infection in Section 4.

2 Methodology

The Dartmouth College campus is relatively compact, with over 190 buildings on about 200 acres. During the period of this study, there were about 566 802.11b access points (APs) covering most of the campus's indoor and outdoor areas. These APs share the same SSID (Service Set IDentifier) to allow seamless roaming between APs. A building's APs, however, are connected to the building's existing subnet, and clients roaming between subnets are forced to obtain new IP addresses.

We installed 18 network "sniffers" in 14 different buildings to obtain detailed network-level traces. The 18 sniffers covered 121 APs in popular places, including libraries, dormitories, academic departments, and social areas. Each sniffer was a Linux box with two Ethernet interfaces, one for data sniffing and one for remote access. In each of the 18 switchrooms we attached the APs to a switch that has a port set to "monitor" mode and was hooked up with the sniffer. We then used tcpdump[1] to capture any wireless traffic going through these APs' wired interfaces (see citehenderson:mobicom04 for further details).

We are mainly interested in studying two peer-to-peer applications, Kazaa[2] and iTunes.[3] Kazaa is a popular *global* file sharing application, both in terms of number of participants and traffic volume seen. A Kazaa host allows the user to search and download files from other Kazaa hosts on the Internet without a centralized server. On the other hand, iTunes allows a user to discover another *local* iTunes host on the same subnet, load its playlist, and stream the audio of selected songs. A Kazaa or iTunes host is then both a server and a client.

In our study we used a four-week tcpdump trace collected by all 18 sniffers during the Winter 2004 term, from February 1 to February 28 inclusive. Though our sniffers had only partial coverage of the campus, we saw 4,084 wireless devices in the tcpdump trace out of all 5,321 wireless devices on campus during the same period (we used another trace to get all wireless MAC addresses [8]). We filtered out Kazaa and iTunes traffic using their respective port numbers in the IP headers. Note that we discarded the iTunes' local peer discovery traffic, since that protocol is used by many other applications. We then derived social networks for Kazaa and iTunes, in terms of undirected graphs, with a node representing a host and an edge representing communications seen in the trace.

There are two limitations of our approach. First, we missed any traffic between two wireless clients associated with the same AP since that traffic does not go through the AP's wired interface and thus was not captured by tcpdump. This implies that the derived social networks had *underestimated* size in terms of the number of nodes and edges. Second, it was difficult to identify unique nodes because a host may change its IP address during the period of our study. For Dartmouth wireless hosts, we can use the MAC address in the tcpdump trace to uniquely identify them. For all other hosts, however, we had to distinguish them using their IP addresses. Unlike the previous case, this limitation implies that we *overestimated* the number of network nodes and edges. Although these factors do not necessarily cancel each other, we believe that they have little impact on our results given the statistical nature of our study.

3 Structural Analysis

We represent social networks as graphs, with nodes representing hosts and edges representing application-level communications. We focus the analysis on the wireless hosts since the social networks we captured were not complete. For instance, we could not observe the links between wired hosts (at Dartmouth or another site). We are also interested in the structural evolution of these social networks, so for some metrics we plot their values in daily sequence on the *accumulated* networks from the first day of our trace. We present the results here and discuss their implications in the next section.

3.1 Network size

Figure 1a shows the evolution of the accumulated overall network size (Y axis in log scale). We observed 22,935 Kazaa participants in total, out of which 329 were Dartmouth wireless hosts and less than 900 were Dartmouth wired hosts. The local iTunes network had 1,114 participants, out of which 616 were wireless hosts. The plot shows that for both networks, the number of new participants decreased in general and started to converge after some period of time. In addition, we also measured for each day how many *new*, previously unseen in the trace, *wireless* hosts participated in the global Kazaa and local iTunes networks. For Kazaa, the number of daily new wireless participants decreased from about 25 to 6 by the end of our trace. For iTunes, the daily new wireless hosts joining the network decreased from 51 to about 13 after 4 weeks.

Figure 1b plots the daily number of wireless participants. For Kazaa, there were maximum 52, minimum 18, and average 31.9 wireless users every day. For iTunes, we saw maximum 113, minimum 47, and average 82.9 daily wireless participants. Both networks had "weekend" effects, though interestingly, the number of wireless iTunes users

[1] http://www.tcpdump.org/
[2] http://www.kazaa.com/
[3] http://www.apple.com/itunes/

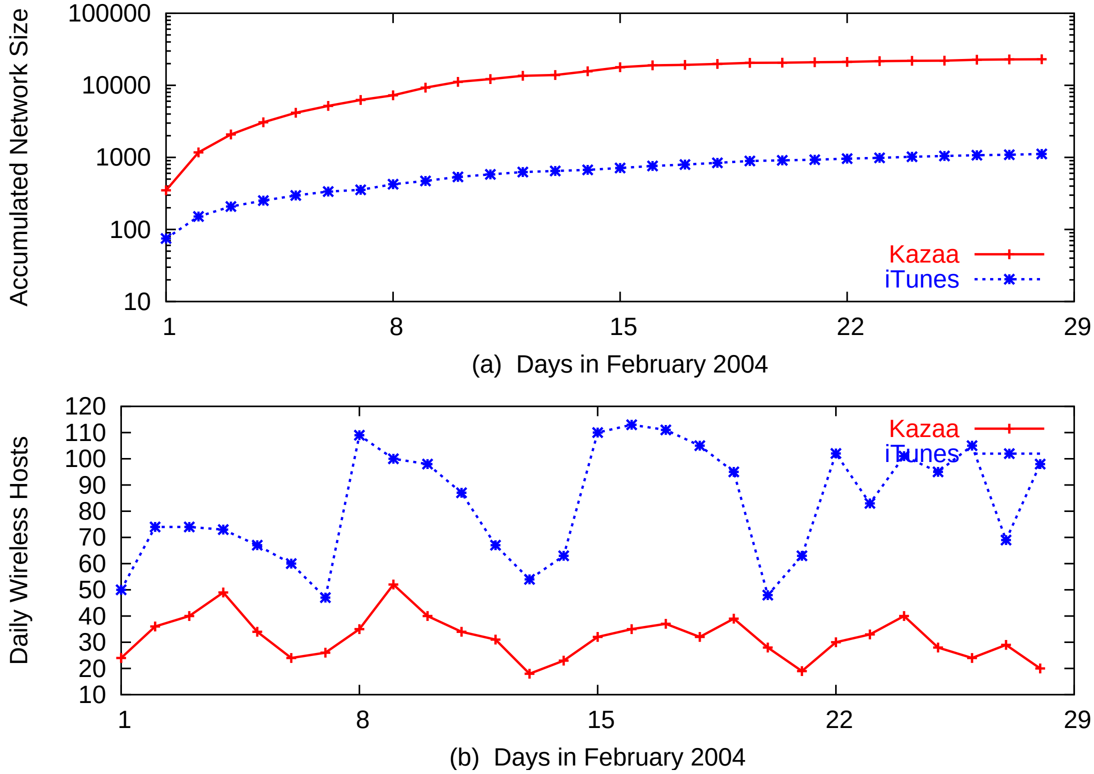

Figure 1: The accumulated network size and the number of daily *wireless* participants. February 1, 2004 was a Sunday.

bounced back on Sundays. One possibility for this could be that many students came back to work on their homework assignments that may be due on Monday, and they listened music while studying.

It is interesting to note that Kazaa appeared to be less popular than iTunes among wireless users. This is a bit counterintuitive given that a large fraction of Internet traffic is due to peer-to-peer applications. It is hard to confirm whether users tend not to use Kazaa wirelessly because of the limited bandwidth. So we looked at what operating systems these wireless hosts used (see [8] for details). We found that 40.3% of wireless iTunes hosts ran Mac OS and 59.3% ran Windows. On the other hand, 89.7% of wireless Kazaa clients ran Windows and only 9.1% ran Mac OS. Given that we had a large pool of wireless Mac users (about half of wireless Windows population) and most Kazaa clients only use Windows, it is then likely that we saw less wireless Kazaa participants than iTunes that supports both Windows and Mac OS.

Next we measured how *active* these wireless clients were in the two social networks. By active, we mean the participant had at least communicated with one peer. Of the wireless Kazaa participants in our 4-week trace, about 56.7% were active only for one day and about 95% were active less than 10 days. It was 46.6% and 89.5% percent for wireless iTunes participants, respectively. There are several possible reasons for the low activity of many wireless participants. First, due to mobility some users may only occasionally show up in the area covered by our sniffers. Second, a poor experience (failing to find what is desired for the first time, or the file turned out not to what desired after long download period) may cause the user to be reluctant to use Kazaa again. Finally, if users are more interested in their own collections instead of shared ones, then they will not often be active on iTunes network.

3.2 Network components

Both Kazaa and iTunes networks were not fully connected graphs and consisted of multiple *components*, where a component is a subgraph in which there is a path between any two nodes. We call the component with maximum number of nodes in the graph the *giant* component. Figure 2 shows the evolution of the networks; subplot (a) gives the number of components and subplot (b) gives the size of the giant component in the networks.

We found few interactions between wireless Kazaa clients, only 34 hosts communicated with another wireless

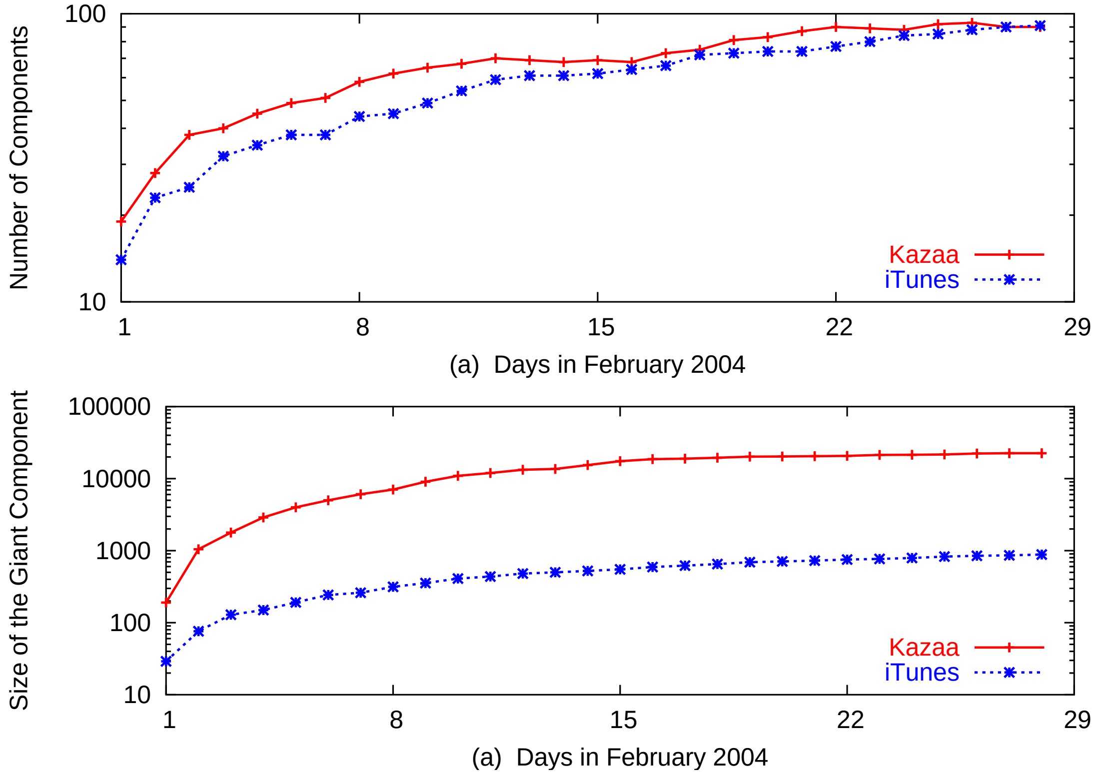

Figure 2: The evolution of the number of components and the size of the giant component.

peer. Thus our Kazaa network was mostly a bipartite graph, with edges connecting wireless and wired hosts. The 22,606 wired Kazaa nodes included fewer than 900 Dartmouth hosts. The continued increase of the number of the components (from 19 to 90) indicates that the neighbor sets of wireless nodes are mostly disjoint. On the other hand, the giant component grew from 190 to 22,567 nodes, covering 98.4% of all Kazaa nodes by the end of the trace. We observed that the giant component already covered almost 90% nodes by the second day. The giant component contained several active Kazaa client(s) contacting many peers for downloading files; for instance, one wireless client communicated with 9,091 peers in total. So, although there was one huge component, there were always many tiny components as well.

It is surprising to see that the much smaller iTunes network having similar number of components as Kazaa network did, growing from 14 to 91. The giant component of iTunes network covered 79.2% of the nodes by the end of the trace, growing from 29 to 882. There are two contributing reasons for the high fragmentation of the iTunes network. First, two iTunes hosts can only communicate when they are on the same subnet, so the network thus have separate components for different subnets and these components may only merge as the hosts move across subnets. Second, the higher number of components (than the number of subnets we monitor) indicate that there were several components in individual subnets due to disjoint user interests. If we consider only the 616 wireless iTunes nodes, however, 425 of them formed a wireless-only network whose nodes communicated with at least one wireless peer. For this sub-network, the number of components actually decreased from 17 to 5 by the end of the trace. The size of the giant component in this wireless-only network increased from 33 to 411. These wireless hosts formed the core of our iTunes network and they became better connected due to host mobility and local interactions.

3.3 Node degrees

We define *degree* of a node to be the number of the edges with it as an end, or equivalently the number of its neighbors. We define *average degree* to be the sum of node degrees divided by the number of nodes. We found that the average degree of the Kazaa network changed little over time and stayed around 2.2. We believe that this low degree was caused by star-shaped components, and the average degree for a star is close to 2 (that is $\frac{2(n-1)}{n}$). The average degree

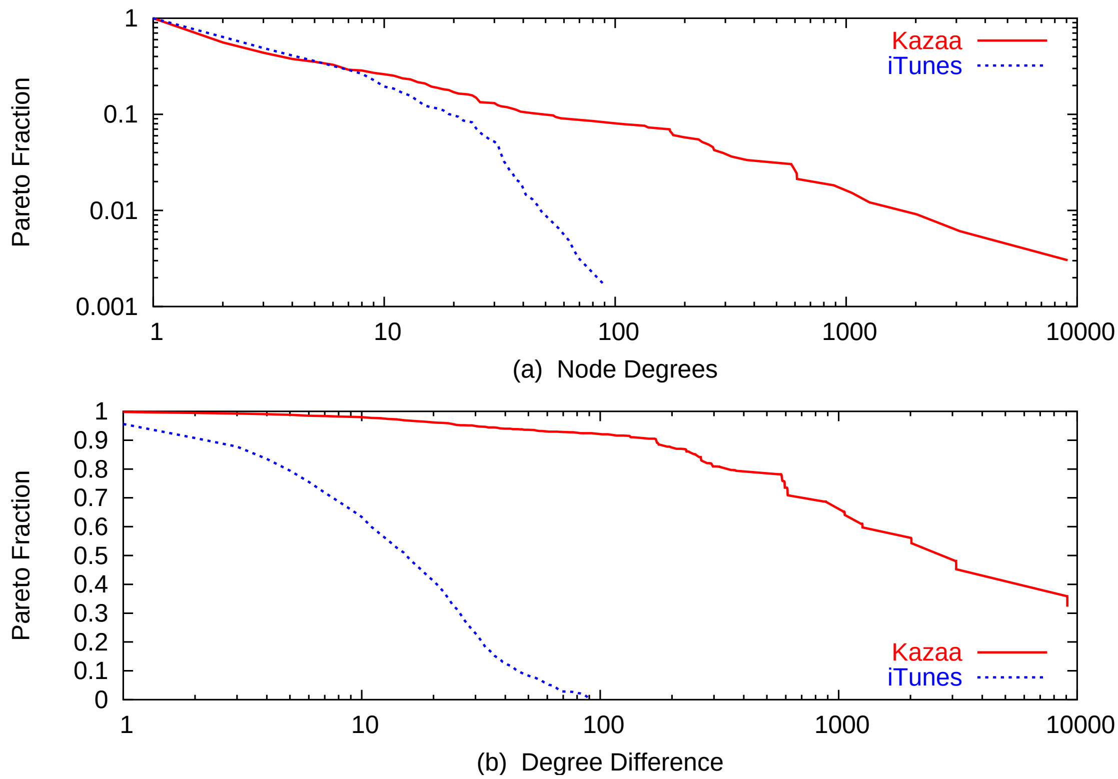

Figure 3: The distributions of node degree and the degree difference of two end nodes of edges.

of the iTunes network, however, increased from 2.1 to 4.8 over the one-month period as the new interactions linked existing nodes. If we consider the wireless-only iTunes sub-network as described previously, the average degree actually increased from 1.9 to 7.0 showing there was abundant edges in the core components.

Figure 3a plots the distribution of the degrees of *wireless* hosts of two networks. The y axis is the *pareto* value, or the fraction of nodes whose degree is greater or equal to x. Note that both x and y axis are in log scale. The Kazaa curve is almost a straight line, indicating their degrees followed a "power-law" distribution [4]. We saw 91.3% Kazaa nodes had degree 1, most of them are wired hosts since we could not observe their other links. A wireless node had the largest degree 9,091, and about 91% of its neighbors had degree 1. There were only 25 other Kazaa nodes with degree more than 100. The degree distribution for iTunes, however, deviated from the power law and was not heavy tailed. About 63.8% iTunes hosts had more than 1 peer, 40.9% had more than 3 peers, 9.4% had more than 20 peers, and the largest degree was 91. Note that there is no "super" iTunes host emerging from the network due to the constraints of local communications.

Figure 3b plots the distribution of degree difference of an edge's two end nodes over all edges. The medians of the degree differences for the Kazaa and iTunes networks are 2,018 and 15 respectively. What it means is that the edges in both networks tended to connect nodes with rather different degrees. As a matter of fact, the average degree of a node's neighbors decreased as that node's own degree increased. This is called *disasortative mixing* in social networks [10]. While our Kazaa and iTunes networks are not globally complete, since we could not observe the links between wired nodes, we believe that both networks had such phenomena and the probability of connecting two high-degree nodes was relatively low.

3.4 Small world

The term "small world" originated with the seminal experiment, conducted by Stanley Milgram in 1967, to test the hypothesis that individuals in United States would be connected through short chain of social ties [14]. A network with the small-world phenomenon is loosely characterized by a short *average path length* and a high *clustering coefficient* [15]. To handle disconnected graphs, we define average inverse path length P to be: $\frac{1}{M}\sum\frac{1}{L}$, the sum of the

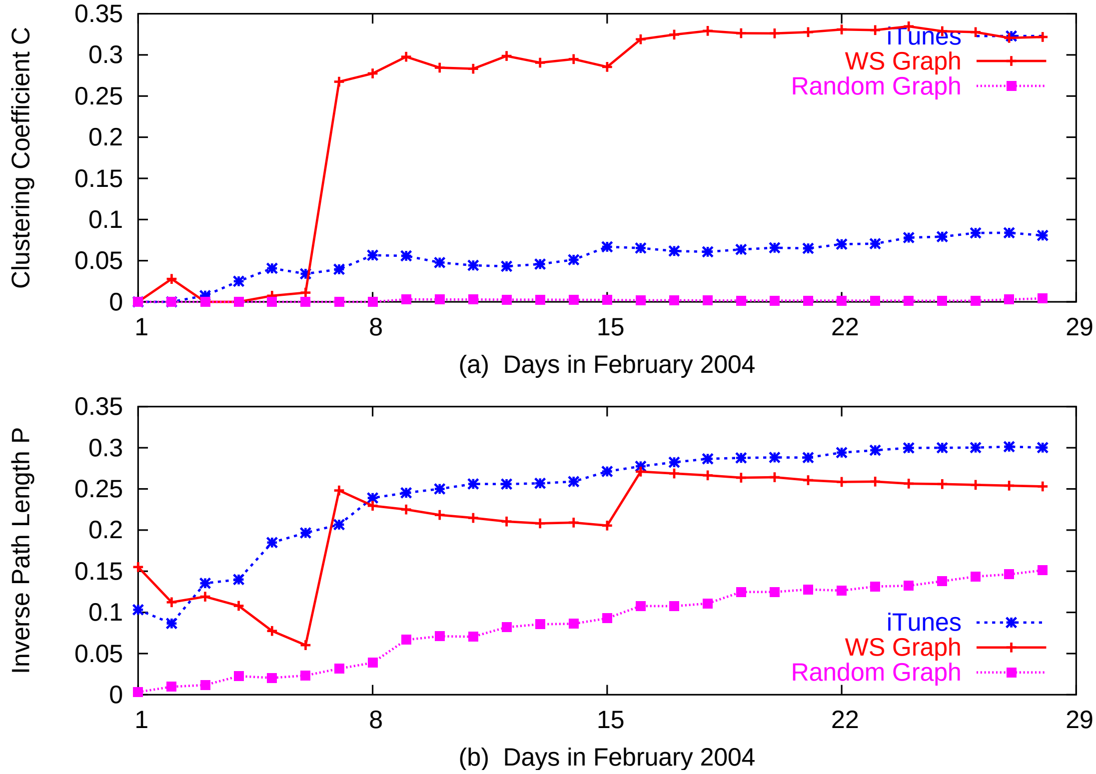

Figure 4: The small world characteristics.

inverse length ($\frac{1}{L}$) of shortest paths between any two nodes, divided by the number of all possible edges in the graph ($M = \frac{N(N-1)}{2}$ and N is the number of nodes). If there is no path between two nodes, the corresponding $\frac{1}{L}$ is zero. Furthermore, assuming C_v is the number of edges among node v's neighbors (size k_v) divided by the number of all possible edges among them ($\frac{k_v(k_v-1)}{2}$), then the clustering coefficient C is defined to be the average of C_v over all nodes. We expect to see both relatively high P and C values for small-world graphs.

In Figure 4 we plot the changing values of these metrics for the iTunes network. We ignore the Kazaa network here since (as we observed it) it was mostly a bipartite graph and did not have small-world characteristics. For comparison, we plot the two metric values for "WS graphs" that was created with an algorithm designed to show small-world characteristics [15]. We used WS graphs with the same number of nodes and approximately same number of edges as the accumulated iTunes networks had on a daily basis. Creating a WS graph requires an even integer for average degree as a parameter, so the result is an approximation to the iTunes network. We also compared the iTunes network with random graphs [2], which take the number of nodes and edges as parameters, so it had same size (nodes and edges) as the iTunes counterparts.

From Figure 4a we saw WS graph's C value had large increases on February 7th when the average degree increased from 2 to 4. With smaller average degrees, even the WS graphs had almost no clustering due to their lattice structure [15]. Beyond February 6th, the WS graphs had high clustering because it had nodes whose neighbors were also well connected. Random graphs had almost no clustering, while iTunes had limited clustering. Examining the inverse path length P in Figure 4b, the WS graphs had a repeated pattern that once the average degree was fixed, the P continued to decrease as the number of nodes increased until next degree increase. This effect is understandable since a lower P implicates longer paths in the graph. The iTunes networks had P values comparable with WS graphs but much higher than random graphs. This result indicates that the iTunes network had a lower average path length as the network evolved. In summary, the iTunes networks had weak small-world characteristics, with short average path lengths and some low-level clustering.

4 Discussion

A recent study of Kazaa network by analyzing traffic collected at Internet backbone shows that the Kazaa node degree does not obey a perfect power-law distribution [12]. Our study, however, shows at a smaller scale that the degree distribution for the wireless nodes follows a power law. We also found that the wireless nodes tend to connect to a large number of off-campus peers for file downloading, while there is little communication among local wireless peers. This result suggests that Kazaa is *locality unaware*, meaning it is likely not the case that Kazaa tries to resolved a query with data already available within an organization. Gummadi et al. propose a distributed caching mechanism to redirect all requests to local peers when possible to reduce external bandwidth consumption [7]. Our analysis results show that this or similar method should also consider redirecting the requests to wired peers or even aggressively replicate the files across peers for improved availability, since we saw generally low wireless Kazaa participation (average 32 daily and 56.7% nodes were only active for one day during 4-week trace).

The iTunes client does not have a query interface for remote peers without after mounting their playlists. This factor is discouraging, since mounting a playlist takes time in the order of tens of seconds. We believe that returning the availability of queried songs and their last seen location will be a valuable service to users, and has potential to enhance social networking of the iTunes community together with other services such as user ratings and a chat interface (not available in iTunes). Supporting such queries without a centralized server, however, would require each node to actively exchange playlists of their own and those they have seen. Eventually each node should know the information of every song in the network through opportunistic playlist exchange. Our analysis results show that this approach will suffer from a slow convergence time since we still had 5 separate components even for the wireless-only sub-network after 4 weeks. However, this approach may be a good approximation since the giant component in the wireless-only sub-network covered almost 97% of participants in our case.

Finally, we are also studying how Internet worms may propagate through these peer-to-peer social networks among wireless hosts, which are often less closely administered and more vulnerable. P2P worms has the potential to directly infect a victim's neighbors since a node often caches its peers' addresses, without the need to scan the Internet to find targets, as do existing worms such as CodeRed [13]. Such non-scanning worms are much stealthier and could dodge existing detection methods [13]. Our results show that if a wireless host is infected in Kazaa network, P2P worms may migrate quickly away from a local site by following edges to remote peers since there is little communication among local Kazaa hosts. This behavior calls for inter-organization cooperation to observe the global patterns for quick response. Unfortunately if the P2P worm could also infect iTunes nodes, then it will travel quickly in the community through established paths in social networks, given that the iTunes network tend to have short paths as we found. We are further investigating this research direction.

5 Summary

We analyzed two Internet-based social networks formed from online interactions, with a focus on wireless participants. The Kazaa users form a global file sharing network and the iTunes users form a local music streaming network. We present our findings regarding several structural properties and their evolution, and we discuss how these results could be leveraged for the design and implementation of wireless applications and supporting systems. As future work, we plan to further study the traffic characteristics in these networks, and to model Internet worms in the wireless social networks using the results presented here.

Acknowledgement

We thank Tristan Henderson for his work on the data-collection infrastructure of the wireless networks at Dartmouth College. We also thank Tristan Henderson, Minkyong Kim, and Bob Gray for their constructive comments on the early draft of this paper. This work is supported by the Institute for Information Infrastructure Protection at Dartmouth College under award #2003-TK-TX-0003 from the U.S. Department of Homeland Security. Points of view in this document are those of the author(s) and do not necessarily represent the official position of the U.S. Department of Homeland Security.

References

[1] Lada A. Adamic, Rajan M. Lukose, Amit R. Puniyani, and Bernardo A. Huberman. Search in power-law networks. *Physical Review*, 64, 2001.

[2] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for massive graphs. In *ACM STOC*, pages 170–180, Portland, OR, May 2000.

[3] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406:378–382, July 2000.

[4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.

[5] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. *Computer Networks*, 33(1-6):309–320, June 2000.

[6] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the Internet topology. In *ACM SIGCOMM*, pages 251–262, Cambridge, MA, August 1999.

[7] Krishna P. Gummadi, Richard J. Dunn, Stefan Saroiu, Steven D. Gribble, Henry M. Levy, and John Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *ACM SOSP*, pages 314–329, Bolton Landing, NY, October 2003.

[8] Tristan Henderson, David Kotz, and Ilya Abyzov. The changing usage of a mature campus-wide wireless network. In *ACM MOBICOM*, pages 187–201, Philadelphia, PA, September 2004.

[9] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *ACM STOC*, pages 163–170, Portland, OR, May 2000.

[10] M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(28), November 2002.

[11] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, April 2001.

[12] Subhabrata Sen and Jia Wang. Analyzing peer-to-peer traffic across large networks. *IEEE/ACM Transactions on Networking*, 12(2), April 2004.

[13] Stuart Staniford, Vern Paxson, and Nicholas Weaver. How to Own the Internet in Your Spare Time. In *Usenix Security*, San Francisco, CA, August 2002.

[14] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[15] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ’small-world’ networks. *Nature*, 393:440–442, June 1998.