Cluster Structures in Topology of Large-Scale Social Networks Revealed by Traffic Data

Masaki Aida*, Keisuke Ishibashi†, Chisa Takano‡, Hiroyoshi Miwa§, Kaori Muranaka† and Akira Miura¶

*Faculty of System Design, Tokyo Metropolitan University
Email: aida@m.ieice.org
†NTT Information Sharing Platform Laboratories, NTT Corporation
‡Traffic Engineering Business Unit, NTT Advanced Technology Corporation
§Department of Informatics, School of Science and Technology, Kwansei Gakuin University
¶Network Laboratories, NTT DoCoMo Inc.

Abstract—Many studies of social networks have recently been published. Interest in topological structures, such as scale-free characteristics, has been particularly strong. In this paper, we focus on the analysis of macro traffic data in a communications network of cellular phone users as a way of investigating large-scale social networks. Behaviors of information exchange between pairs of cellular phone users are reflected in traffic data, which thus reflects interesting features of social networks. We analyze the relationship between the number of customers and the volume of traffic with a view to finding clues about the structure of social networks among the very large set of potential customers. We then demonstrate some interesting features that our analysis reveals: a scale-free topology of human relations, their cluster structures, and behaviors of user-dynamics. In addition, we consider the relationship between traffic volume and the number of customers depending on the situation.

I. INTRODUCTION

The number of customers for NTT DoCoMo’s ‘i-mode service’ in Japan (for Internet access from cellular phones) had been increasing explosively and reached about ten million within one and half years of service introduction in February 1999 [1]. Statistics on i-mode’s growth thus provide an interesting body of information about the behavior of large numbers of users. Complex networks such as structures of social relationships do not have an engineered architecture; rather, they are self-organized by the actions of large numbers of individuals. Local interactions can lead to the nontrivial global phenomenon of a scale-free distribution of node degree [4], which in turn leads to a small-world property [2], [3]. In this paper, we analyze the relationship between the number of i-mode customers and the volume of traffic, with a view to finding clues about the structure of human relations among the very large set of potential i-mode customers. The traffic data reveals that this structure is a scale-free network, and we calculate the exponent that governs the distribution of node degree in this network. The data also indicates that people who have more friends tend to subscribe to the i-mode service at an earlier stage.

If we consider each person as a node and each relationship between two people as a link, we have a graphical model of human relations. If we can systematically characterize the structure of graphs thus derived, the characterization should be applicable to marketing strategies, the propagation of rumors and epidemics, and demand forecasting for telecommunications services, among others.

Networks of hyperlinks among web pages on the Internet and certain social networks have been reported to show small-world properties and act as scale-free networks. A scale-free network has a small number of ‘hub’ nodes, each of which has a very large number of links. This feature acts to suppress increases in network diameter when the number of nodes increases. As a result, the average numbers of hops in the routes between all pairs of nodes are extremely small, and information spreads with remarkable speed. The defining feature of a scale-free network is that the distribution of node degree obeys a power law, i.e., \( n(k) \propto k^{-\gamma} \), where \( k \) and \( n(k) \) denote the node degree (number of links) and the number of nodes of degree \( k \), respectively, and \( \gamma \) is a positive constant. A wide variety of scale-free networks has been found, in both the technological and social realms. In most cases, the relation \( 2.0 \leq \gamma \leq 3.4 \) applies [5]. Ebel et al. [6] analyzed the logs of the e-mail server at Kiel University and produced a graph that represents the relationships among the e-mail accounts of the students. A link in the graph indicates the passage of at least one e-mail message between the corresponding pair of accounts. The graph in this case was a scale-free network with the slightly atypical value of 1.81. This result reflects human relations within a small community, in this case the set of people who use the university’s e-mail server. The result is thus not applicable to people in general. Furthermore, since an e-mail log will almost certainly include records of multicast messages, i.e., messages sent to accounts on a mailing list, the passage of an e-mail message does not necessarily indicate a relationship between the owners of the corresponding pair of accounts, so the result does not solely reflect human relations. Abello et al. [7] and Aiello et al. [8] analyzed telephone calls on a certain day and produced a graph that represents the relationships among phone numbers. In this case, a link represents the setting up of a connection between the corresponding pair of numbers. This graph is a scale-free network with \( \gamma = 2.1 \). The data in this case concerns a large number of unspecified people, so the result should be generally applicable; however, a phone number often corresponds to a...
company or family rather than an individual, so the result does not reliably reflect human relations.

In [9], we investigated the simple relationship between the number of i-mode customers and the volume of their traffic and derived the structure of human relations for the large set of potential customers of the service. In our previous work, we did not take into consideration the fact that the connectivity of human relations is changed by the length of traffic observation period. Two customers are linked if and only if they have at least one message exchange. So, if we observe traffic data for a longer period, customers are densely connected by links and it may change the observed structures of human relation. In addition, the degree of the scale-free network that describes human relations was inappropriately shown in our previous work. In this paper, we take the length of the traffic observation period into consideration, and we investigate the change in the structure of human relations and the cluster structure that appears in them.

The rest of this paper is organized as follows. In section II, we introduce our traffic data and point out the characteristics that make them desirable as a basis for analyzing human relations. In section III, we explain our assumptions and the framework of our investigation and present the analytical and general form of the relationship that characterizes i-mode e-mail usage. In sections IV and V, we assume stochastic and deterministic rules, respectively, for the selection of new i-mode subscribers and investigate the patterns of human relations they reflect and the user dynamics we would expect if the rule were correct. A deterministic rule described in Sec. V reveals that the structure of human relations forms a scale-free network. In section VI, the results obtained from Sec. V are extended to the structure of human relations observed during an observation period of finite length. In addition, we combine the structure with a cluster structure. In section VII, we show case studies about the relationship between the behavior of the observed volume of traffic and service’s penetration process. Finally, we conclude our discussion in section VIII.

II. TRAFFIC DATA

Data about i-mode service traffic is of particular interest for the following reasons.

(a) The explosive growth of the service minimizes the effect on traffic of external factors such as changes in economic circumstances, family structure, and so on.

(b) Since almost all cellular phone terminals are for individual use, the transfer of an e-mail message between two terminals unambiguously represents communication between the corresponding pair of customers.

(c) Almost all e-mail traffic in the service is one to one, so we can assume that the volume of e-mail traffic is proportional to the number of customer pairs exchanging messages.

(d) Sending an e-mail message is much cheaper than making a voice call, so external factors, e.g., the income of users, only have negligible effects on the traffic patterns.

(e) In the early stages of popularization of the i-mode service, the combination of few e-mail advertisements and little sensationalism to attract nuisance users meant that very little of the traffic was independent of relationships among people.

The number of customers grew about three fold, from 1,290,000 to 3,740,000, over the six months from Aug. 1999 through Jan. 2000 [1]. The relationship between the number of customers and the volume of i-mode web-service traffic in this period is shown in Fig. 1. Let the number of i-mode customers be \( m \); the relationship is then written as

\[ \text{(i-mode web traffic)} \propto m. \]  

The most reasonable explanation for this is a stable frequency of web access per user. The reason for this is as follows: if users who have subscribed to the i-mode service at an earlier stage are heavier users than more recent subscribers, the volume of web traffic will not be proportional to \( m \). The result thus implies that the usage characteristics of the i-mode service for the average customer were stable over this period.

The number of i-mode customers and the number of i-mode messages in the same period is given in Fig. 2. The data follows this power law:

\[ \text{(i-mode e-mail traffic)} \propto m^{1.55}. \]  

Therefore, the number of e-mail messages increased more quickly than the volume of web traffic. If the number of e-mail messages for an average customer is independent of the number of i-mode customers and is stable, it should be proportional to \( m \). A value greater than \( m \) reflects growth over time in the number of partners with whom the average customer might want to communicate. On the other hand, if the average customer knows a certain constant proportion of customers (even if this proportion is small, e.g., 0.0001%), the volume of e-mail traffic should be proportional to \( m^2 \). The fact that the volume of e-mail traffic is proportional to \( m^{1+\alpha} \), where \( 0 < \alpha < 1 \), means that while the number of possible communication partners increases, the ratio of this number to the number of all customers falls. The parameter \( \alpha = 0.55 \) characterizes the rate of growth for e-mail traffic. It also tells us something about the strength of human relations. Hereafter, we investigate the characteristics of human relations that satisfy (2).

III. NOTATION AND ASSUMPTIONS

Let the set of people in Japan (i.e., the set of all potential customers of the i-mode service) be \( V \), and the set of pairs of people \((x, y) \in V \times V\) who exchange information with each other be \( E \). The number of elements in \( V \) is \(|V| = n\). We define human relations as a graph \( G(V, E) \). We assume that \( G(V, E) \) is stationary.

Let the set of pairs of people \((x, y) \in V \times V\) who exchange information with each other when we observe the traffic for time period \( \tau \) be \( E(\tau) \) \((E(\tau) \subset E)\), and human relations
when we observe the traffic for time period \( \tau \) be \( G(V, E(\tau)) \). \( G(V, E(\tau)) \) is a subgraph of \( G(V, E) \) and includes the same nodes in \( G(V, E) \). \( G(V, E) = \lim_{\tau \to \infty} G(V, E(\tau)) \).

Next, we consider the set of i-mode customers \( V_m (V_m \subset V) \). The number of elements in \( V_m \) is \( |V_m| = m \) (\( m \leq n \)). Let the subgraph induced by \( V_m \) from \( G(V, E) \) be \( G_m(V_m, E_m) \). \( G_m(V_m, E_m) \) represents human relations among i-mode customers.

Let the set of pairs of people \((x, y) \in V_m \times V_m\) who exchange information with each other when we observe the traffic for time period \( \tau \) be \( E_m(\tau) \) (where \( E_m(\tau) \subset E_m \)), and human relations among i-mode customers when we observe the traffic for time period \( \tau \) be \( G_m(V_m, E_m(\tau)) \), which is a subgraph of \( G_m(V_m, E_m) \) and includes the same nodes in \( G_m(V_m, E_m) \). Here, \( G_m(V_m, E_m) = \lim_{\tau \to \infty} G_m(V_m, E_m(\tau)) \).

Since the usage characteristics of i-mode service for an average customer are unchanged, for a fixed \( \tau \), we assume that the number of links, \( |E_m(\tau)| \), in the subgraph \( G_m(V_m, E_m(\tau)) \) is proportional to the volume of e-mail traffic (as the number of messages) flowing in the i-mode service\(^1\):

\[
|E_m(\tau)| = O(m^{1 + \alpha}).
\]

To clarify the origin of the behavior that leads to (2), i.e., to \( 0 < \alpha < 1 \), we need to find the conditions of human relations \( G(V, E(\tau)) \), and not just of the subset of relations \( G_m(V_m, E_m(\tau)) \).

In the following two sections, we assume two possible rules for selecting subscribers to the i-mode service, identify the rule that corresponds with our data, that is, the rule that characterizes user-participation dynamics, and show the structure of human relations \( G(V, E(\tau)) \) thus implied.

IV. RANDOM SELECTION OF NEW I-MODE CUSTOMERS

As preparation, we show a case where a new i-mode customer is selected at random from \( V \), and consider the relationship between the structure of human relations and the volume of traffic. The discussion follows ref. [9].

We sort all elements of \( V \) into descending order of degree (number of links connected to the element) with respect to the graph \( G(V, E(\tau)) \) and let the degree of the \( i \)-th element be \( D_i(\tau) \) (\( i = 1, 2, \ldots, n \)). In cases where multiple nodes have the same degree, \( i \) is assigned arbitrarily. Similarly, all elements of \( V_m \) are sorted into descending order of degree with respect to the subgraph \( G_m(V_m, E_m(\tau)) \). We let the degree of the \( j \)-th element be \( d_j(\tau) \) (\( j = 1, 2, \ldots, m \)).

Next, let us consider continuous versions of the degree distributions \( D_i(\tau) \) and \( d_j(\tau) \), denoted by \( D(x; \tau) \) and \( d(y; \tau) \), respectively, where \( 0 \leq x \leq n \) and \( 0 \leq y \leq m \). The distribution \( D(x; \tau) \) of degree is a monotonically decreasing function of \( x \), and we choose \( D(x; \tau) \) that satisfies

\[
\sum_{i=a}^{b} D_i(\tau) = \int_{a-1}^{b} D(x; \tau) \, dx,
\]

where arbitrary parameters \( a \) and \( b \) are integers that satisfy \( 1 \leq a \leq b \leq n \). We choose the distribution \( d(y; \tau) \) in a similar way.

Since the elements of \( V_m \) are chosen from \( V \), a node that has the \( j \)-th largest degree, \( d_j(\tau) \) (\( j \in V_m \)), in \( G_m(V_m, E_m(\tau)) \), will on average correspond to a node in the set with the \( i \)-th (\( i = (n/m)j \)) largest degree in \( G(V, E(\tau)) \). In addition, since the probability of the nodes connected to node \( i \in V \) in \( G(V, E(\tau)) \) being in \( V_m \) is \((m-1)/(n-1)\), we get

\[
d_j(\tau) \simeq \frac{m-1}{n-1} D_i(\tau),
\]

on average.

\(^1\)This assumption was indirectly verified by using traffic of voice communications among cellular phone terminals.
Therefore, the number of links in $G_m(V_m, E_m(\tau))$, that is, $|E_m(\tau)|$, is expressed as
\[
|E_m(\tau)| = \frac{1}{2} \sum_{j=1}^{m} d_j(\tau) = \frac{1}{2} \int_0^{m} d(y; \tau) \, dy \\
\approx \frac{1}{2} \frac{m-1}{n-1} \int_0^{m} D((n/m)y; \tau) \, dy \\
= \frac{1}{2} \frac{m-1}{n-1} \int_0^{n} D(x; \tau) \, dy \, dx \\
= \frac{1}{2} \frac{m(m-1)}{n(n-1)} \int_0^{n} D(x; \tau) \, dx \\
= \frac{1}{2} \frac{m(m-1)}{n(n-1)} \sum_{i=1}^{n} D_i(\tau) \\
= O(m^2).
\]

Consequently, if we assume that new customers of i-mode are selected at random, the volume of e-mail traffic is independent of the structure of human relations and increases by $O(m^2)$. This does not agree with (3).

V. STRUCTURE OF HUMAN RELATIONS AND RULE FOR I-MODE SERVICE SUBSCRIPTIONS

From the previous section, we can see that the rule for selecting $m$ elements of $V_m$ from $V$ is not a random sampling. In this section, we further investigate the node selection rule by a deterministic approach [9].

We sort all elements of $V$ into descending order of degree (number of links connected to the element) and let the degree of the $i$-th element be $D_i$ ($i = 1, 2, \ldots, n$). In cases where multiple nodes have the same degree, $i$ is assigned arbitrarily. Similarly, all elements of $V_m$ are sorted into descending order of degree in the subgraph $G_m(V_m, E_m)$. We let the degree of the $j$-th element be $d_j$ ($j = 1, 2, \ldots, m$). On the other hand, we sort those elements of $V$ that have been selected for $V_m$ into their order of selection and let the degree within $V$ of the $k$-th element be $D_k^{(s)}$ ($k = 1, 2, \ldots, n$). All elements of $V_m$ are sorted into the same order. Let the degree within $V_m$ of the $h$-th element be $d_h^{(s)}$ ($h = 1, 2, \ldots, m$). While the $g$-th element, $g \in V_m$, corresponds to $g \in V$, $d_g^{(s)}$ is very rarely the same as $D_g^{(s)}$.

We assume that, for an average user, the ratio of those who are i-mode users among friends is proportional to the degree of penetration of i-mode service. This means the average node degree in $G_m(V_m, E_m(\tau))$ is proportional to the degree of penetration of i-mode service. Then, we can choose a function $c(m)$ in
\[
\sum_{h=1}^{m} D_h^{(s)}(\tau) \approx c(m) \sum_{h=1}^{m} D_h^{(s)}(\tau)
\]
which is monotonically decreasing with respect to $h$. The above results are rephrased below.

- The distribution of node degree in $G(V, E(\tau))$ obeys Zipf’s law [10], [11] with a power index $-1 - \alpha$,
\[
D_h(\tau) = O(h^{-(1-\alpha)}).
\]

- The elements of $V_m$ tend to be selected from $V$ in descending order of degree in $G(V, E(\tau))$. In other words, a person who has many friends tends to subscribe to the i-mode service at an earlier stage.

Conversely, the above two results along with $c(m) = O(m)$ lead us to (3).
The defining feature of a scale-free network is that the distribution of node degree obeys a power law,

\[ n(k) \propto k^{-\gamma}, \]  

where \( k \) and \( n(k) \) denote the node degree (number of links) and number of nodes of degree \( k \), respectively, and \( \gamma \) is a positive constant. This relation is called Lotka’s law [11]. From the fact that the distribution of node degree in \( G(V,E(\tau)) \) obeys Zipf’s law, we can show that the human relationship \( G(V,E(\tau)) \) is a scale-free network.

In [9], we showed that

\[ \gamma = \frac{1}{1 - \alpha} \]

from a graphical consideration, but that is not appropriate. The corrected discussion is as follows. If the distribution of node degree \( D_i(\tau) \) obeys Zipf’s law with a power index \(-\beta\) (in this case, \( \beta = 1 - \alpha \)), this can be written as

\[ D_i(\tau) = C \, i^{-\beta}, \]  

where \( C \) is a positive constant. We consider \( i \) and \( j \) that respectively give \( D_i(\tau) = k \), \( D_j(\tau) = k - 1 \). They are represented as

\[ i = C^{1/\beta} \, k^{-1/\beta}, \]

\[ j = C^{1/\beta} \, (k - 1)^{-1/\beta}. \]  

Since \( n(k) \) is given by \( j - i \), we have

\[ n(k) = j - i = C^{1/\beta} \left\{ (k - 1)^{-1/\beta} - k^{-1/\beta} \right\} \]

\[ \simeq C^{1/\beta} \, k^{-1/\beta} \, \frac{1}{\beta k} \]

\[ = O(k^{-(1/\beta + 1)}). \]  

Therefore,

\[ \gamma = \frac{1}{\beta} + 1 = \frac{1}{1 - \alpha} + 1. \]  

Using \( \alpha = 0.55 \), we have \( \gamma = 3.22 \). Consequently, human relationship \( G(V,E(\tau)) \) is a scale-free network with \( \gamma = 3.22 \).

VI. CHANGES OF HUMAN RELATIONS AND THEIR CLUSTER STRUCTURES WITH RESPECT TO OBSERVATION PERIOD LENGTH

The results in the previous section are based on the assumption \( c(m) = O(m) \). If we choose a larger observation period \( \tau >> 1 \), there are many links in \( E_m(\tau) \). In this case, it is natural to assume that, for an average user, the ratio of those who are i-mode users among friends is proportional to the degree of penetration of the i-mode service. However, this assumption may be doubtful for a general \( \tau \).

(3) denotes the relationship between the number of i-mode users and the volume of their traffic observed in a day (\( \tau = 1 \) day). If (3) for \( \tau = 1 \) day, the same relationship is valid for a larger observation period (e.g., \( \tau = \) week, month, ...). This implies that the discussion in the previous section is for \( \tau \to \infty \). The results obtained from the previous section (a scale-free network with \( \gamma = 3.22 \)) are not valid for the human relationship \( G(V,E(\tau)) \) with general \( \tau \), but are restricted to the human relationship \( G(V,E) \) with observation period \( \tau \to \infty \).

For finite \( \tau \), we introduce \( c(m, \tau) \) instead of \( c(m) \), choose \( c(m, \tau) \neq O(m) \), and consider

\[ c(m, \tau) = O(m^{\delta(\tau)}), \]  

where \( \lim_{\tau \to \infty} \delta(\tau) = 1 \). From \( |E_m(\tau)| = O(m^{1+\alpha}) \) and it is independent of \( \tau \), we have

\[ \sum_{h=1}^{m} D_h^{(s)}(\tau) = O(m^{1+\alpha - \delta(\tau)}). \]  

If (21) is valid for all \( m \), we have

\[ D_h^{(s)}(\tau) = O(h^{\alpha - \delta(\tau)}). \]  

If \( \alpha - \delta(\tau) > 0, \) (22) leads unphysical situations that a person who has few friends tends to subscribe to the i-mode service at an earlier stage. If \( \alpha - \delta(\tau) = 0, \) \( \alpha = 1 \) because \( \lim_{\tau \to \infty} \delta(\tau) = 1 \). If \( \alpha - \delta(\tau) < 0, \) we apply the same discussion in the previous section and obtain the following results.

- The distribution of node degree in \( G(V,E(\tau)) \) obeys Zipf’s law with a power index \(-\delta(\tau) - \alpha\),

\[ D_h(\tau) = O(h^{-(\delta(\tau) - \alpha)}). \]  

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2005 proceedings.
• The elements of $V_m$ tend to be selected from $V$ in descending order of degree in $G(V,E(\tau))$.
• human relationship $G(V,E(\tau))$ is a scale-free network with
  \[
  \gamma = \frac{1}{\delta(\tau) - \alpha} + 1.
  \]  

Let us consider the penetration ratio of the i-mode service $\rho = m/n$. The fact that the order of $c(m, \tau) = O(m^{\delta(\tau)}) = O(\rho^{\delta(\tau)})$ ($\delta(\tau) < 1$) implies that, for an average user, the ratio of those who are i-mode users among friends increases faster than the penetration ratio of the i-mode service, which is proportional to $\rho$. This means that there exist sets of users (called clusters), and a new user tends to be selected from the same cluster as the old users. Note that although such cluster structures and the scale-free structures depend on $\tau$, they are mutually related because $|E_m(\tau)| = O(m^{1+\alpha})$ is independent of $\tau$.

VII. CASE STUDIES: STRUCTURE OF HUMAN RELATIONS AND VOLUME OF TRAFFIC

Based on the above results, we consider the relationship between the structure of human relations and the volume of traffic.

a) Case where a New User is Selected at Random and is Independent of the Structure of Human Relations: In this case, $c(m, \tau) = O(m)$ and $\sum_{h=1}^{m} D_h^{(s)}(\tau) = O(m)$. Therefore, we have
  \[
  \sum_{h=1}^{m} d_h^{(s)}(\tau) = O(m^2).
  \]
This means that the observed volume of traffic should be proportional to $\rho^2$.

b) Case where a New User is Selected Independent of Its Node Degree and $c(m, \tau)$ is Independent of $m$: In this case, $c(m, \tau) = O(m^0)$ and $\sum_{h=1}^{m} D_h^{(s)}(\tau) = O(m)$. Therefore, we have
  \[
  \sum_{h=1}^{m} d_h^{(s)}(\tau) = O(m).
  \]
This means that the observed volume of traffic should be proportional to $\rho$. This case includes the following situations.

• There is little traffic among users. That is, they are in different clusters.
• New users are selected from the same clusters.
• New users are selected from the same clusters.

Case where a New User is Selected in Descending Order of Degree and Depends on the Structure of Human Relations: We consider the case where the observed volume of traffic is proportional to $\rho^{1+\alpha}$ ($0 < \alpha < 1$):
  \[
  \sum_{h=1}^{m} d_h^{(s)}(\tau) = O(m^{1+\alpha}).
  \]
In this case,
  \[
  c(m, \tau) = O(m^{\delta(\tau)}), \quad (\delta(\tau) < 1),
  \]
for $\alpha - \delta(\tau) < 0$, means that users with a large node degree are in the same cluster and a new user is selected not only from the same cluster as old users but also from a new cluster.

VIII. CONCLUDING REMARKS

In this paper, we took into consideration the length of the traffic observation period and investigated the change in the structure of human relations and the cluster structure that appears in them. We corrected the value of power index of human relations given in [9]. In addition, by taking the structure of human relations into consideration, we considered the relationship between the behavior of the observed volume of traffic and penetration process of service.

To achieve wide penetration of a new service, two factors are important: both the increase in the number of users that are not restricted to specific clusters and the increase in the number of frequent communication users in specific clusters. The increase in the number of users that is independent of the structure of human relations does not sufficiently cause the increase of the volume of traffic. If we know the structure of human relations and their cluster structure, we should be able to obtain the behavior of the service’s penetration process by observing the behavior of the traffic volume. This technique will enable early judgement of whether a service will succeed.

REFERENCES