

Chapter 2

Structure and Dynamics of Social Networks Revealed by Data Analysis of Actual Communication Services

Masaki Aida and Hideyuki Koto

2.1 Introduction

Up to now, data of actual communication services obtained from communication networks, such as the volume of traffic and the number of users, has mainly been used to forecast traffic demands and provision network facilities. It can be said that this use focuses on the “quantitative” side of the data. On the other hand, such data can also illuminate several characteristics of the structures of the human society. This chapter introduces a new “qualitative” use of communication network data. We try to extract social information from the data, and investigate the universal structure of social networks that underlie the most popular communication services. Our expectation is that each communication service provides a different window on the universal social network structure. The question is how to access those windows.

A direct technique for examining social network structures is the questionnaire approach. However, its extremely high cost makes it impractical if we target a comprehensive analysis of the universal social network structure. Our solution is to collect and analyze the quantitative data generated by communication services. The contents of communication logs etc. offer views of the individual situations associated with that service. Examples of these situations include interpersonal relationships in an organization and agreements reached between corporations. However, conventional analysis fails to extract complete images of the universal social network structure.

In this chapter, we focus on the power laws that appear in the data of actual communication services. In explaining the reasons that underlie the power laws, we elucidate the whole universal social network structure. Since the power laws examined in this chapter describe the relations present in coarse data, detailed behaviors (e.g., who is communicating with who) of each user cannot be observed. However, we can expect to extract a more nearly universal structure that is independent of the superficial structures present in each data set. Once we develop a universal model of

M. Aida (✉)
Tokyo Metropolitan University, Hino-shi, Tokyo 191-0065, Japan
e-mail: maida@sd.tmu.ac.jp

social networks, we can better understand the process of service penetration and can find a better activation method that can replace the word-of-mouth communication-based marketing approach, for not only existing services but also future services. In addition, a comprehensive understanding of the universal social network structure could be applied to not only communication services but also more general commodities and services such as business and marketing strategies.

We explain here why we focus on power laws. We know that certain types of distributions (e.g., normal, Poisson, etc.) originate from randomness. Differing from these distributions, power laws can be assumed to have deterministic causes. Therefore, investigation of the reason of power laws is not disturbed by randomized effect, and the cause of power laws is connected to other phenomena.

Our approach is summarized as follows. We analyze three different data sets: the volume of traffic in the initial stage of NTT DoCoMo's i-mode service [3], the logs of NTT DoCoMo's voice traffic, and the number of mixi users [9]. Hereafter, we call these data sets as Service I, II, and III, respectively. Service I is the first Internet access service offered over cellular phone terminals, Service II is a cellular phone service, and Service III is the largest social networking service (SNS) in Japan. By combining these analyses we obtain three results with regard to the social networks that underlie specific communication services. The first is the degree distribution of social networks, the second is the topological rules of social networks, and the last is user dynamics with regard to the actions needed to join a communication service. The first result was verified through a cross-check using different data; the logs of voice traffic presented by KDDI's cellular phone service [11]. We call this data set as Service IV.

The rest of this chapter is organized as follows: Section 2.2 provides a conceptual image of the methods available for analyzing social networks. Section 2.3 analyzes, according to [1, 2], the data of the cellular phone service (Service I and II) to derive partial information on the social network structure. The partial information so obtained cannot completely determine the model of social networks and there is an undetermined parameter in the model. Section 2.4 analyzes data on SNS (Service III) users to supplement the partial information obtained in Sect. 2.3. The combined use of both results enables us to determine the value of the parameter in the social network model. The result is a social network model that is self-consistent with the data observed from different services (Service I, II, and III). In Sect. 2.5, we verify the validity of our social network model by using the traffic logs of a cellular phone service that were not analyzed in earlier sections (Service IV). Section 2.6 concludes our discussion with a brief summary.

2.2 Analysis Strategy

We use graph $G(V, E)$ to represent the relationship of people exchanging information, where V is a set of nodes (people) and E is a set of links (information exchanges) between nodes. We call $G(V, E)$ the social network.

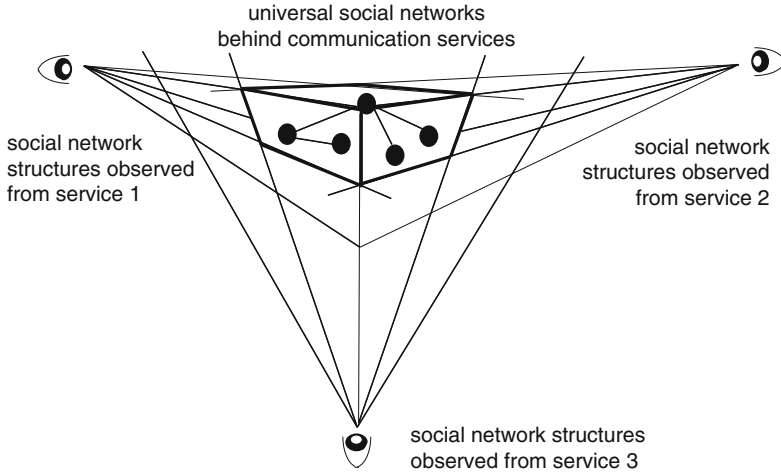


Fig. 2.1 The relationship between the universal social network and images obtained from specific communication services

The global structure of $G(V, E)$ cannot, unfortunately, be observed directly although the object of our interest is to clarify the structure of $G(V, E)$. Our solution is to adopt the approach of investigating the structure of $G(V, E)$ indirectly; we analyze specific communication services, such as cellular phone and SNS. Our purpose is not to investigate the specific services themselves, but to use them to elucidate the structure of social network $G(V, E)$.

How then is it possible to extract the universal social network structure? The concept of our approach is illustrated in Fig. 2.1. The network at the center of Fig. 2.1 is the “multi-dimensional” social network, and the three eyes represent three different services that hold partial information of the social network as “contracted” information. Although the “universal” social network at the center of the figure cannot be observed directly, we assume that sets of partial information can be extracted from specific communication services. These partial information sets may allow us to construct the “multi-dimensional” or “universal” social network model by combining them.

2.3 Analysis of Social Networks Based on Traffic Data of Internet Access Service Offered Over Cellular Phones

In this section, we introduce the partial information set created by analyzing the data of the cellular phone service.

2.3.1 *Data To Be Analyzed*

This subsection analyzes the data that holds the relationship between the number of users and email traffic during the early growth period of Service I; the world's first Internet access service from cellular phone terminals [3]. Since Service I was launched on February 22, 1999, the service has seen an explosive increase in the number of users. In the first one and half years (up to August 2000) the number of users exceeded ten million. The process by which a network service can acquire users at such a dramatic rate offers an interesting window on the structure of social networks and user behavior regarding hot-selling products.

This set of Service I data is useful for understanding social networks because it has the following properties:

- Since the number of Service I users increased explosively within a short period, it can be assumed that the Service I traffic was little affected by external factors such as a change in people's lifestyle.
- Since most cellular phones are exclusively used by their owners, traffic between cellular phones can be regarded as information exchange between people.
- Since most Service I emails are one-to-one communication, it can be assumed that email traffic is closely related to the number of pairs of Service I users who are exchanging information with each other.
- Since the cost of sending an email is far lower than that of talking on the phone, it can be assumed that the volume of email communication is little affected by such external factors as the income level of the individual users.
- Since the early period of the Service I had few problems with unwanted advertising emails sent to users indiscriminately, it can be assumed that almost all traffic arose from existing social networks.

During the early expansion period, 6 months from the beginning of August 1999 to the end of January 2000, the number of Service I users increased almost threefold, from 1,290,000 to 3,740,000. The relationship between the Web traffic (number of Web access attempts) and the number of Service I users during this period can be modeled as:

$$(\text{Web traffic}) \propto n, \quad (2.1)$$

where n is the number of users (chart on the left in Fig. 2.2). This is self-evident as long as the average number of Web access attempts per user is constant. Conversely, the fact that the above relation holds means that people's average usage of the Service I service did not change during this period. In other words, there is no evidence that the earliest subscribers to the Service I were heavier users. Meanwhile, the volume of email traffic (number of email messages) can be modeled as:

$$(\text{Email traffic}) \propto n^{5/3}. \quad (2.2)$$

Thus, a power law applies (chart on the right of Fig. 2.2). If the volume of communication per user remained constant even as n increased, then the volume of email traffic should be proportional to n . The fact that email traffic is proportional to $n^{1+\alpha}$

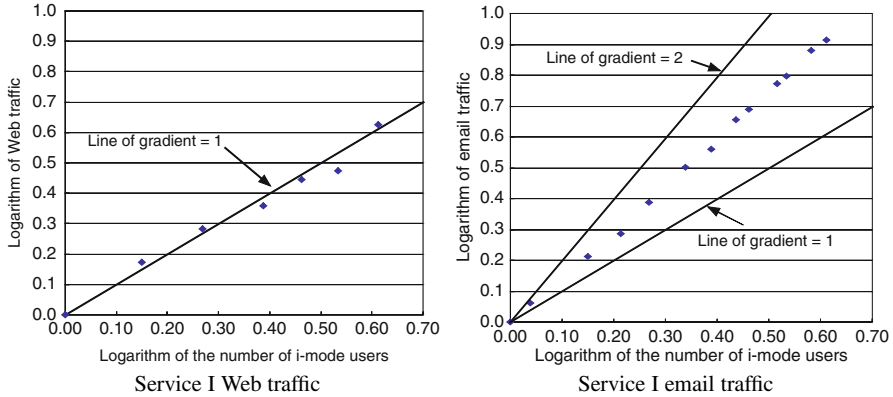


Fig. 2.2 The relationship between the number of users and the volume of Service I traffic for Web and email

($\alpha \simeq 2/3$) suggests that an increase in n results in an increase in the number of Service I users a single user communicates with. Therefore, $\alpha \simeq 2/3$ characterizes the rate of increase in email traffic. This also tells us something about the strength of human relations in social networks.

The following examines the graphical structure of universal social networks $G(V, E)$, involving not only Service I users but also others, using the power law (2.2) identified from the email traffic data described above.

2.3.2 Definition of Symbols and Problem Description

As mentioned in Sect. 2.2, $G(V, E)$ represents the social network, and the number of people in V is N ($|V| = N$). We assume that $G(V, E)$ does not change over time.

We use a rule to select n nodes from V ; the subset of these selected nodes is $V_i(n)$ ($n \leq N$). Let $G_i(V_i(n), E_i(n))$ be the subgraph induced by $V_i(n)$ from $G(V, E)$. That is, a node pair is connected by a link in $G_i(V_i(n), E_i(n))$ if and only if the corresponding node pair in $G(V, E)$ is connected by a link. Each element of $V_i(n)$ is an Service I customer and social networks among all Service I customers are represented by $G_i(V_i(n), E_i(n))$ (see Fig. 2.3).

Equation (2.1) indicates that the usage of Service I by individual users did not change even as the number of Service I users increased. Therefore, it can be assumed that the traffic per link between a user and a Web site remained constant. Similarly, we assume that the average email traffic per link is also constant irrespective of the number of Service I users.¹ Thus, the number of links $|E_i(n)|$ becomes,

$$|E_i(n)| = O(n^{1+\alpha}). \quad (2.3)$$

¹ The fact that traffic per link is not affected by the number of Service I users, n , has been indirectly confirmed from Service II. See Appendix A for details.

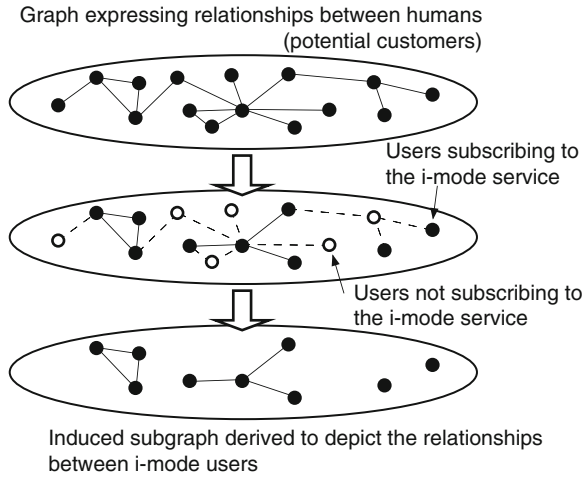


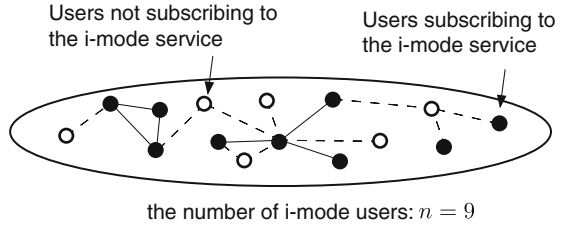
Fig. 2.3 Example of $G(V, E)$, a graph showing the structure of the social networks, and $G_i(V_i(n), E_i(n))$, the subgraph induced from Service I users

The issue addressed by this paper is not the study of $G_i(V_i(n), E_i(n))$, or social networks established between Service I users, but $G(V, E)$, or universal social networks among both users and non-users of the Service I, as indicated by the traffic data of Service I. Figure 2.3 shows the relation between $G(V, E)$ and $G_i(V_i(n), E_i(n))$. The upper graph, $G(V, E)$, shows universal social networks while the bottom graph is a subgraph, $G_i(V_i(n), E_i(n))$, derived from $G(V, E)$, showing the social networks among Service I users. The number of Service I users and the volume of email traffic correspond to the number of nodes and the number of links, as derived in (2.3), in $G_i(V_i(n), E_i(n))$. The structure of $G(V, E)$ and how people begin to subscribe to the Service I are considered below.

2.3.3 How People Subscribed to the Service I and the Structure of Social Networks

First, we introduce two different schemes for numbering the elements of V , and define three sequences of node degree (the number of links that a node has) based on the numbering.

We call the node with the largest node degree as node 1. Similarly, we call the node with the j th largest node degree as node j . In addition, let the magnitude of node degree of node j be $D(j)$. Next, we introduce another numbering of elements in V according to the time of subscribing to the Service I. Let $D_i(\ell)$ be the node degree of the ℓ th earliest subscribed node in $G(V, E)$. Similarly, let $d_i(n, \ell)$ be the degree of the ℓ th earliest subscribed node with respect to $G_i(V_i(n), E_i(n))$ when the number of Service I users is n .

Fig. 2.4 Example of $c_i(n)$ 

Assume that the degree of Service I user in $G_i(V_i(n), E_i(n))$ can be related to his or her degree in $G(V, E)$ as follows:

$$\sum_{\ell=1}^n d_i(n, \ell) = c_i(n) \sum_{\ell=1}^n D_i(\ell), \quad (2.4)$$

where $c_i(n)$ indicates the ratio of the number of Service I user's acquaintances subscribing to the Service I to the total number of acquaintances, given that the number of Service I users is n . That is

$$c_i(n) = \frac{2 \times (\text{total number of links between Service I users})}{\text{total number of Service I users' degrees w.r.t. } G(V, E)}.$$

The function $c_i(n)$ is a monotonically increasing function with $c_i(1) = 0$ and $c_i(N) = 1$. Figure 2.4 shows an example of $c_i(n)$. In this case, $N = 15$, $n = 9$, and

$$\sum_{\ell=1}^n D_i(\ell) = 22, \quad \sum_{\ell=1}^n d_i(n, \ell) = 12, \quad c_i(n) = \frac{6}{11}.$$

We assume the following power function as a property of $c(n)$:

$$c_i(n) \propto n^{1-\delta}, \quad (2.5)$$

where δ is a constant. The validity of the assumption (2.5) is discussed below.

Since $c_i(n)$ will increase as the penetration of the Service I increases, δ should satisfy $\delta < 1$. Here it is worth to note the relationship between the value of δ and topology of the social networks.

- If $\delta > 0$, since $c_i(n)$ is convex, this inequality indicates that $c_i(n)$ grows rapidly in the early stage of the Service I. In other words, there is something about cluster structures in that earlier subscribers to the Service I are more likely to be acquaintances of each other. If $\delta = 0$, this means that there is no evidence of the above cluster structures. Otherwise, $\delta < 0$ is not realistic because this would mean that later subscribers of the Service I were more likely to be acquaintances of each other.

From (2.3) and (2.4), we can derive

$$\sum_{\ell=1}^n D_i(\ell) \propto n^{\alpha+\delta}, \quad (n \ll N). \quad (2.6)$$

If this holds for any n of $n \ll N$, then

$$D_i(\ell) \propto \ell^{\alpha+\delta-1}, \quad (\ell \ll N). \quad (2.7)$$

Here, let us consider three cases identified by the value of $\alpha + \delta - 1$. First, in the case of $\alpha + \delta - 1 < 0$, $D_i(\ell)$ decreases with respect to ℓ . Therefore, $D_i(\ell)$ is the node degree of the ℓ th earliest subscribed node in $G(V, E)$, and it is simultaneously the node degree of the ℓ th largest magnitude of node degree. This correspondence is not so strict but is valid for accuracy in terms of observations in logarithmic charts. Consequently, if $\alpha + \delta - 1 < 0$, we have

$$D_i(\ell) \simeq D(\ell) \quad (\text{in terms of order}), \quad (2.8)$$

for $\ell \ll N$. This relation leads to the following results.

- The node degree of social networks $G(V, E)$ obeys Zipf's law where the exponent is $-(1 - \alpha - \delta)$,

$$D(\ell) \propto \ell^{-(1-\alpha-\delta)}, \quad (\ell \ll N). \quad (2.9)$$

- People tend to subscribe to the Service I in the order of decreasing degree in $G(V, E)$. In other words, people with more acquaintances tend to subscribe to the service earlier.

This finding about the who subscribed to the Service I service first can be considered to mirror the tendency generally cited in the marketing area where people with higher sensitivity to information (more acquaintances) are more likely to try something before it becomes known or popular.

Next, in the case of $\alpha + \delta - 1 = 0$, $D_i(\ell)$ is independent of ℓ . It is known that if we construct an induced subgraph by selecting nodes in $G(V, E)$ at random, the number of links in the induced subgraph is proportional to n^2 where the number of selected nodes is n [1]. This is independent of the structure of $G(V, E)$, and means $\alpha = 1$. From (2.2), the number of links should be proportional to $n^{1+\alpha}$ ($\alpha \simeq 2/3$). Therefore, the assumption of $\alpha + \delta - 1 = 0$ contradicts the observed data of the actual service.

Finally, in the case of $\alpha + \delta - 1 > 0$, people tend to subscribe to the Service I in the order of increasing degree in $G(V, E)$. In other words, people with fewer acquaintances tend to subscribe to the service earlier. This result contradicts our personal experience. From the above considerations, we regard the assumption $\alpha + \delta - 1 < 0$ as being valid.

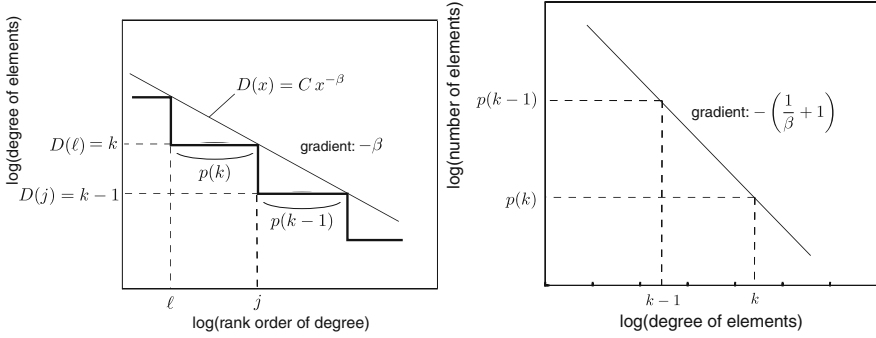


Fig. 2.5 Two points are extracted from data that satisfies Zipf's law (*left*), and they are plotted to give the distribution of degree (*right*)

If the distribution of the degree of nodes in the graph representing social networks follows Zipf's law, social networks can be taken as being scale-free. A scale-free network is a graph in which the distribution of the degree of the nodes follows a power law [6, 7],

$$p(k) \propto k^{-\gamma}, \quad (2.10)$$

where k is the degree of a node, $p(k)$ is the number of nodes with degree k , and $\gamma > 0$ is a constant.

Assume that $D(\ell)$ follows a Zipf distribution with a gradient of $-\beta$ (where $\beta = 1 - \alpha - \delta > 0$) as

$$D(\ell) = C \ell^{-\beta}, \quad (2.11)$$

where C is a constant. Consider ℓ and j for which $D(\ell) = k$ and $D(j) = k - 1$, then

$$\ell = C^{1/\beta} k^{-1/\beta}, \quad j = C^{1/\beta} (k - 1)^{-1/\beta}. \quad (2.12)$$

Since $p(k)$ is $j - \ell$ when $D(\ell) = k$,

$$\begin{aligned} p(k) &= C^{1/\beta} \left\{ (k - 1)^{-1/\beta} - k^{-1/\beta} \right\} \\ &\simeq C^{1/\beta} k^{-1/\beta} \frac{1}{\beta k} \\ &= O(k^{-(1/\beta+1)}). \end{aligned} \quad (2.13)$$

Hence, the graph representing social networks is a scale-free graph whose exponent, γ , is

$$\gamma = \frac{1}{\beta} + 1 = \frac{1}{1 - \alpha - \delta} + 1. \quad (2.14)$$

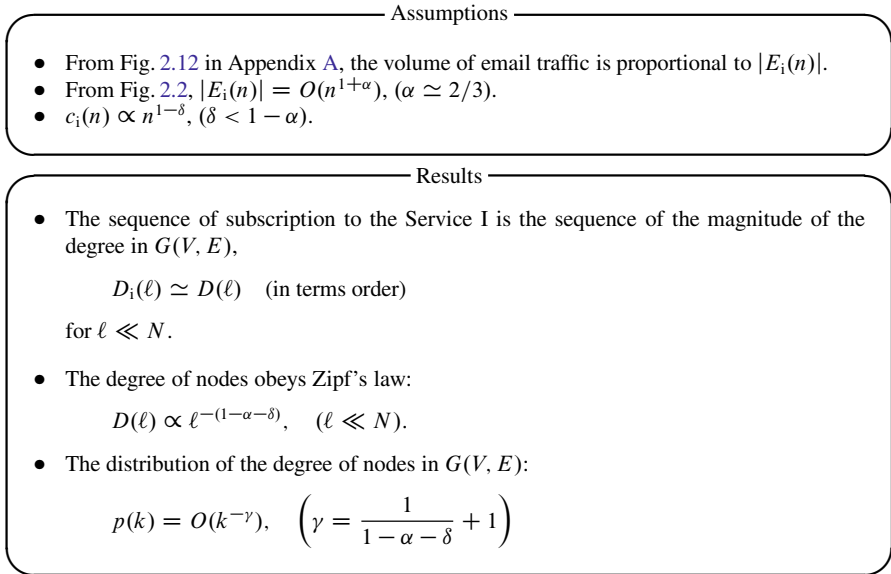


Fig. 2.6 The assumptions made in the analysis of service I data and the results

The assumptions made in the above discussion and its results are summarized in Fig. 2.6.

Although the above discussion does not lead to a specific value for δ , δ should satisfy $\alpha + \delta - 1 < 0$. The fact that $\alpha + \delta < 1$ indeed holds will be supported along with the assumption of $c_i(n)$ in the next section through an analysis of the Service III.

2.4 Analysis of Social Networks Based on the Number of SNS Users

In this section, we investigate the structure of social networks $G(V, E)$ from a different viewpoint, i.e., data generated by the Service III. In addition, by combining these results with the results of our analysis of Service I data, we clarify the details of the social network model including the verification of our assumption of the power law of $c_i(n)$ and the determination of the value of γ .

2.4.1 Analyzed Data

Service III is Japan's largest social networking service provided by mixi, Inc. [9].

For a person to become a member of Service III, he or she needs to be invited to join by an existing member. Although this is a system where only those invited by

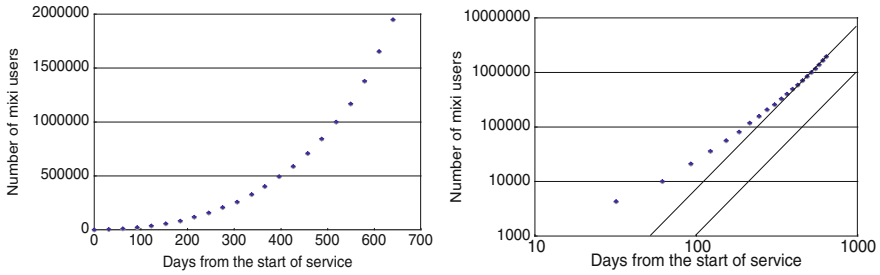


Fig. 2.7 Growth in the number of service III users

existing members may join, the number of users is growing rapidly due to the fact that the mechanism of existing members inviting new people to join is functioning well. The Service III started in February 2004. The number of users reached one million on August 1, 2005, and two million on December 6, 2005. While it took 17.5 months for the number of users to reach one million, it took only 4 months for the number to grow by another million. Because of the following characteristics of the growth in the number of users, Service III data is useful for understanding social networks.

- Since the number of Service III users grew explosively over a short period of time, it can be assumed that the process behind its growth was little affected by external factors such as changes in peoples' lifestyles.
- Since a person needs to be invited to join by an existing member, the process of the growth process of its popularity, i.e., number of users, is closely related to links in social networks.

The left chart in Figure 2.7 shows the growth in the number of Service III users in the early days of the service after launch with 600 users. The horizontal axis is the number of days elapsed since the start of the service. The vertical axis is the number of Service III users. The right chart is a double logarithmic chart. The lines with the gradient of 3 are shown for reference. It was reported that the number of Service III users grew exponentially [10]. Excluding the very early days, when the growth depended on the initial conditions, it can be seen that the growth in the number of users was time to the power of three.

2.4.2 Growth in the Number of SNS Users and Social Networks

Let $m(t)$ be the number of Service III users at time t , and assume that the following holds:

$$m(t) \propto t^3. \quad (2.15)$$

Then, the rate of growth in the number of Service III users, dm/dt , can be expressed as

$$\frac{dm}{dt} \propto t^2. \quad (2.16)$$

Substituting t in (2.15) into (2.16), we get

$$\frac{dm}{dt} \propto m^{2/3}. \quad (2.17)$$

Next, we consider the degree of Service III users. We assume that the potential users of the Service III service are the same as those of the Service I, i.e., the set of V . In other words, the target customers (including potential customers) are the same for both services, specifically, the targets are people living in Japan. We sort the elements in V according to the sequence of the time of joining the Service III service, and let $D_x(\ell)$ be the degree of the ℓ th element in $G(V, E)$. Let $d_x(m, \ell)$ be the degree of the ℓ th element with respect to the graph consisting of Service III users alone (See the middle graph in Fig. 2.8). As in the case of the Service I, function $c_x(m)$ is introduced to relate $d_x(m, \ell)$ to $D_x(\ell)$ as follows:

$$\sum_{\ell=1}^m d_x(m, \ell) = c_x(m) \sum_{\ell=1}^m D_x(\ell). \quad (2.18)$$

$c_x(m)$ indicates the ratio of the number of Service III users' acquaintances subscribing to the Service III to the total number of his or her acquaintances, when the number of Service III users is m , that is

$$c_x(m) = \frac{2 \times (\text{total number of links between Service III users})}{\text{total number of Service III users' degrees w.r.t. } G(V, E)}.$$

It is a monotonically increasing function with $c_x(1) = 0$ and $c_x(N) = 1$.

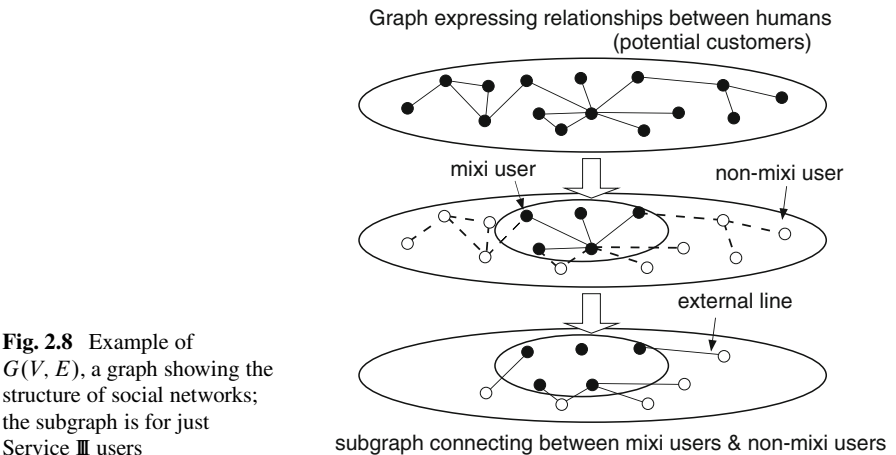


Fig. 2.8 Example of $G(V, E)$, a graph showing the structure of social networks; the subgraph is for just Service III users

The upper and middle figures of Fig. 2.8 show social networks $G(V, E)$ and the induced subgraph of $G(V, E)$ (just Service III users), respectively. The links that are connected to Service III users but do not interconnect Service III users in $G(V, E)$ are referred to as external lines (see the figure at the bottom of Fig. 2.8). There are six such lines in this example. The number of external lines can be expressed as

$$\sum_{\ell=1}^m D_x(\ell) - \sum_{\ell=1}^m d_x(m, \ell) = (1 - c_x(m)) \sum_{\ell=1}^m D_x(\ell).$$

Since the expansion of the Service III service depends on the invitations made by existing members, it is reasonable to assume that the rate of growth in the number of Service III users is proportional to the number of external lines. In other words,

$$\frac{dm}{dt} \propto (1 - c_x(m)) \sum_{\ell=1}^m D_x(\ell). \quad (2.19)$$

Therefore, from (2.17)

$$(1 - c_x(m)) \sum_{\ell=1}^m D_x(\ell) \propto m^{2/3}. \quad (2.20)$$

Thus, a major difference in analyses made on the Service I and III services is the relations being analyzed. The analysis of Service I considered the relations between Service I users, while that of Service III considers the relations between Service III users and non-Service III users. The analysis of the Service I itself cannot illuminate the details of $c_i(n)$, since it only analyzes the relations between users. On the other hand, since the analysis of Service III service considers relations between users and non-users, $c_x(m)$ appears in the form of $(1 - c_x(m))$ in (2.20). The analysis of Service III, therefore, provides different view of social networks than that of Service I.

Let us consider the region of m such that $c_x(m) \ll 1$. For m in this region, the following holds

$$(1 - c_x(m)) \simeq \text{constant}, \quad (c_x(m) \ll 1) \quad (2.21)$$

on a log scale². We can then extract the behavior of the degree from (2.20) as

$$\sum_{\ell=1}^m D_x(\ell) \propto m^{2/3}, \quad (m \ll N), \quad (2.22)$$

where $m \ll N$ means $c_x(m) \ll 1$ since $c_x(m)$ is an increasing function of m . If the extracted behavior of the degree is true for any value of m where $m \ll N$, we get

$$D_x(\ell) \propto \ell^{-1/3}, \quad (\ell \ll N). \quad (2.23)$$

² The validity of (2.21) is verified in Appendix B.

Therefore, the order of subscribing to the Service III follows the order of the magnitude of the degree,

$$D_x(\ell) \simeq D(\ell), \quad (\text{in terms of order}), \quad (2.24)$$

for $\ell \ll N$ and we have

$$D(\ell) \propto \ell^{-1/3}, \quad (\ell \ll N). \quad (2.25)$$

The characteristics of $D(\ell)$ obtained from the analyses of Services I, II, and III should be the same since the targets of both analyses are the same social networks $G(V, E)$. Therefore, by comparing (2.25) and (2.9), we find that

$$\alpha + \delta \simeq \frac{2}{3}, \quad (2.26)$$

for $\ell \ll N$. Using $\alpha \simeq 2/3$, we have the following results:

- For $n \ll N$, $c_i(n)$ is a power function expressed as (2.5).
- The value of δ , which could not be determined by the analysis of the Service I alone, can be determined as

$$\delta \simeq 0. \quad (2.27)$$

Therefore, $c_i(n)$ is a linear function of n .

- δ is such that the inequality, $\alpha + \delta - 1 < 0$, holds.

Moreover, functions $c_i(n)$ and $c_x(m)$ have the same meaning; If we select n nodes (or m nodes) in order of the magnitude of node degree and construct the induced subgraph, both functions represent the ratio of the total number of node degrees in the induced subgraph to the total number of node degrees of selected nodes. Therefore, $c_x(m) \propto m$, and (2.21) is valid for $m \ll N$.

In addition, as mentioned in Sect. 2.3.3, if $\delta > 0$, there are cluster structures, in which earlier subscribers to the Service I are more likely to be acquaintances of each other. The result $\delta \simeq 0$ means that such cluster structures are not observed.

By considering the analyses of the Services I, II, and III data, we can summarize the properties of the structure of social networks that satisfy both Services I, II, and III data as follows. They identify a self-consistent model of social networks obtained from different communication services.

- In general, social networks can be expressed as scale-free graphs with degree distribution of $p(k) \propto k^{-\gamma}$, where γ is

$$\gamma = \frac{1}{1 - \alpha - \delta} + 1 \simeq 4. \quad (2.28)$$

That is, the degree distribution of social networks is

$$p(k) \propto k^{-4}. \quad (2.29)$$

- Function $c_i(n)$, which is an indication of the strength of a cluster of n people sorted according to the magnitude of their degrees, is given by

$$c_i(n) \propto n^{1-\delta} \simeq n. \quad (2.30)$$

This means that there are no cluster structures and $c_i(n)$ is proportional to the penetration ratio n/N (i.e., proportional to the number of users n) of Services I.

The assumptions of the above discussion and its results are summarized in Fig. 2.9.

The above properties are useful in constructing a network model that replicates the characteristics of social networks. Using the constructed network model, we can simulate various processes regarding the penetration of communication services, the mechanism of word-of-mouth communication, and various marketing strategies.

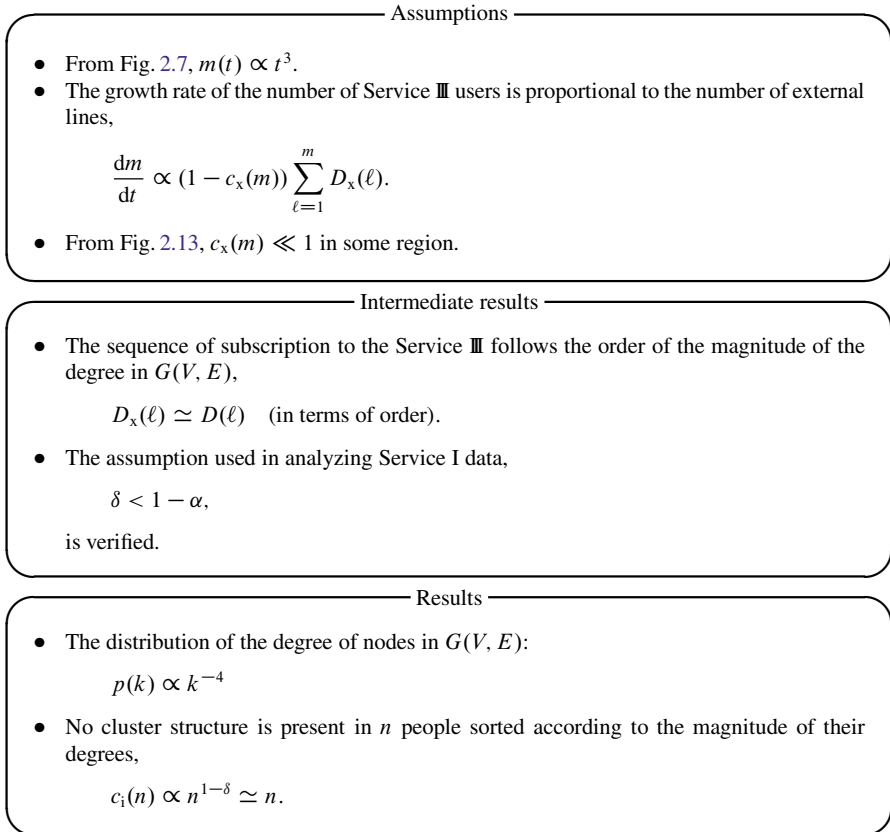


Fig. 2.9 The assumptions made for analyzing data and the results

2.5 Verification of Degree Distribution of Social Networks

If the social network structure obtained in the previous section is universal and is independent of characteristics of specific services (e.g., Services I, II, and III), the obtained structure should be validated by data of another communication service. In this section, we verify the degree distribution of social networks by using the logs of cellular phone traffic, i.e., data that was not used in the aforementioned analyses.

We collected the logs of Service IV, the voice communication service of a cellular phone network. The procedures used for validation are as follows. First, we constructed a graph describing the social networks linking Service IV users by using Service IV's log data. The method used to construct the graph was simple. A node denotes a user, and two nodes are connected by a link if and only if there is communication between the users in some observation period. Links are differentiated for incoming and outgoing calls so the graph is a directed graph. Link means there is at least one call, i.e., the link remains the same regardless of the number of calls in excess of one. In addition, the link is independent of call holding time. We analyzed the graph describing social networks of Service IV users and investigated the distribution of node degrees evidenced by the graph.

Note that we can only investigate the Service IV users in a certain sub-area of the service. In other words, the analysed data and results are not the universal, or total, social network $G(V, E)$. To verify the social network model by using this data, it is necessary that the graph obtained from Service IV data has the same characteristics as the universal social network $G(V, E)$. In general, it is known that the characteristics of the distribution of node degree are the same in both graphs if the users are selected independently of their node degree. It is natural to assume that the subset of Service IV users to be analyzed is selected independent of the node degree, and the selection of service area to be analyzed is also independent of the node degree. Therefore, if the probability distribution of node degree is $p(k)$, the degree distribution of Service IV users in a certain sub-area of the service is also $p(k)$.

The data analyzed here are logs of voice communication over a cellular phone service at six different switches. We analyzed 12 h logs and counted the number of incoming and outgoing calls for each user ID. The number of unique people calling the user is the node degree of incoming calls. The number of unique people called by the user is the node degree of outgoing calls.

Figures 2.10 and 2.11 show the distribution of node degrees (PDF) of the outgoing and incoming calls for each area, respectively. The horizontal axis denotes degree k and the vertical axis denotes the probability density $p(k)$ for degree k , in log scales. We can recognize that the tail distributions are proportional to k^{-4} . These results verify the scale-free property (2.29) of social networks $G(V, E)$.

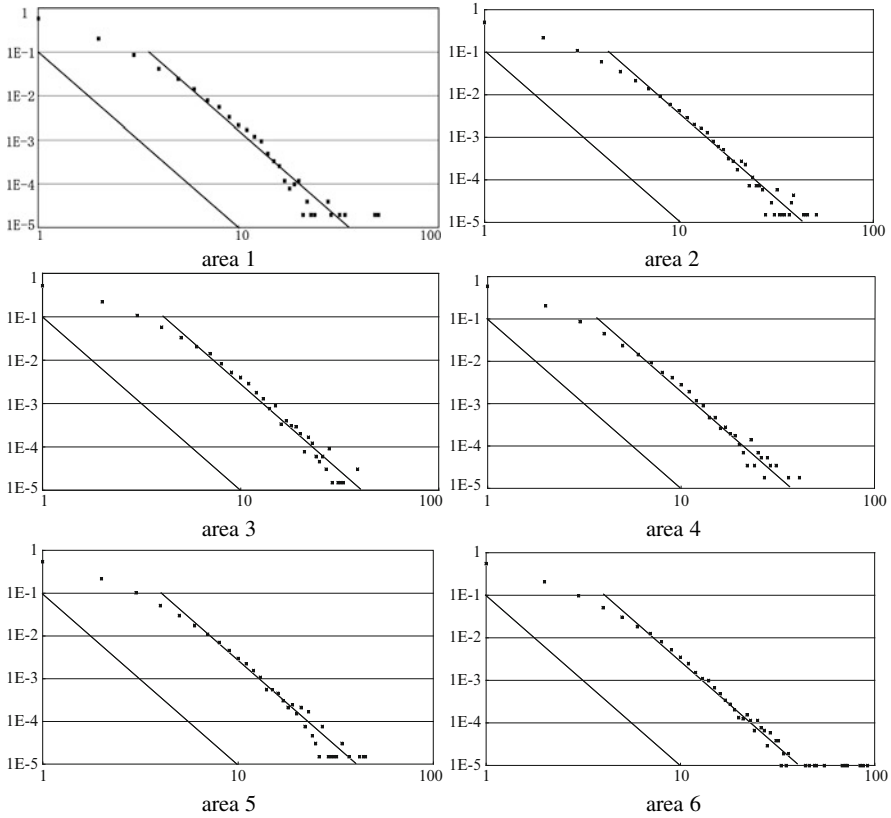


Fig. 2.10 The degree distributions of outgoing calls

2.6 Conclusions

This chapter extracted social information from the data generated by specific communication services, and investigated the universal structure of social networks that underlie each services. In addition, the structure of social networks was verified. A key point of this research is its use of coarse data for analyzing social networks. The analyses examined the relationship between the volume of traffic and the number of users, and the temporal evolution of the number of users of SNS service. These data do not, of course, describe the behaviors of each user. However, we found the structure of social networks as characterized by the distribution of node degree, topological structure of social networks, and user dynamics.

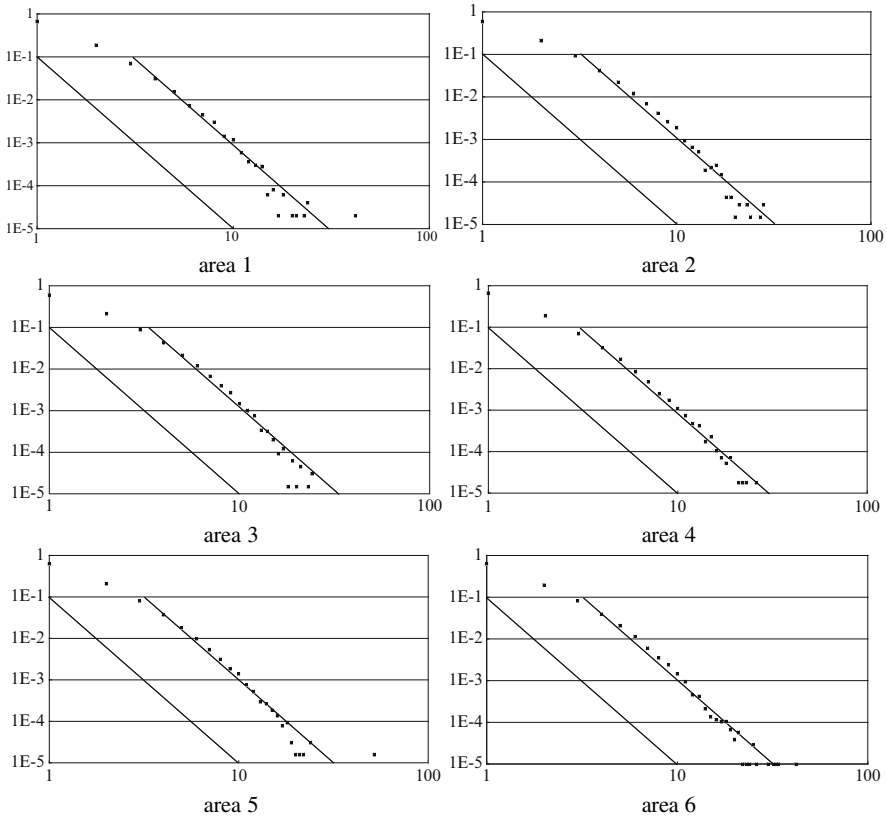


Fig. 2.11 The degree distributions of incoming calls

The features of this work can be summarized as follows.

- Our purpose in analyzing the data of an actual communication service was to extract the structure of the universal social network that is behind the services, not to determine the characteristics of the services themselves.
- The data gathered from a single communication service provides only partial information of the universal social network. By combining the results of data analysis from different communication services, we can extract the detailed structure of social networks.
- We can derive node structures by fitting the data to power laws; note that the data does not describe the detailed behaviors of individual users.
- We cannot verify the analysis results by experiments because we target large-scale social networks. However, we can verify the validity of data analysis from different communication data generated by services that share the common social network.

The characteristics of the social network obtained by the analysis are not characteristics unique to any specific communication service but are universal.

In fact, the obtained model is self-consistent with characteristics of different communication services. Therefore, our findings on social network structure makes it possible to design and engineer some approaches to encourage the penetration of new communication services and information marketing strategies. For example, to improve the speed of the spread when new communication service is introduced, the selecting method of initial users has been studied [12].

Acknowledgements A part of this research was made possible by funds provided by the International Communication Foundation (ICF) (now KDDI Foundation) in its Research Support Program for fiscal year 2005, and by the Grant-in-Aid for Scientific Research (S) No. 18100001 (2006–2010) from the Japan Society for the Promotion of Science.

References

1. M. Aida, K. Ishibashi, H. Miwa, C. Takano, and S. Kuribayashi, "Structure of human relations and user-dynamics revealed by traffic data," *IEICE Transactions on Information and Systems*, vol. E87-D, no. 6, pp. 1454–1460, 2004
2. M. Aida, K. Ishibashi, C. Takano, H. Miwa, K. Muranaka, and A. Miura, "Cluster structures in topology of large-scale social networks revealed by traffic data," *IEEE GLOBECOM 2005*, St. Louis, 2005
3. DoCoMo Net, How the i-mode service is used. <http://www.nttdocomo.co.jp/>
4. Masaki Aida, Jun Sasaki, "Structural analysis on social networks using the spread process of communication services," *IEICE Tech. Rep.*, IN2006-41, vol. 106, no. 151, pp. 37–42, 2006 (in Japanese)
5. R. Rousseau, "*George Kingsley Zipf: life, idea, his law and informetrics*," *Glottometrics*, vol. 3 (To Honor G.K. Zipf), pp. 11–18, 2002
6. A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999
7. R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, no. 47, 2002
8. Masaki Aida, "Structures of social networks and user-dynamics revealed by power laws, Inspired from phenomenology," *Journal of IEICE*, vol. 91, no. 10, pp. 891–896, 2008 (in Japanese)
9. mixi, Inc. <http://mixi.co.jp/>
10. K. Yuda, N. Ono, and Y. Fujiwara, "Human network structure in a social networking service," *Transactions of the Information Processing Society of Japan*, vol. 47, no. 3, pp. 865–874, 2006 (in Japanese)
11. au by KDDI, <http://www.au.kddi.com/>
12. T. Hirano, M. Uwajima, C. Takano, and M. Aida, "Spreading strategy of communication service based on a social network model," *IEICE Tech. Rep.*, IN2008-135, vol. 108, no. 458, pp. 19–24, 2009 (in Japanese)

A Relationship Between the Number of Links and the Volume of Traffic

In order to study the graphic representation of social networks by analyzing Service I traffic, we hypothesized that the observed volume of traffic is proportional to the

number of links in the graph. To verify this hypothesis, it is necessary to analyze the communications log data of individual users. The rough data shown in Fig. 2.2 is not sufficient for such verification.

Since detailed communications log data of Service I users during the same period as that used to construct Fig. 2.2 was not available, we attempted to verify the above hypothesis indirectly by using other types of communications log data that were available. We have examined the number of calls between pairs of subscribers (caller ID and callee ID) in the communications log data of a cellular phone voice communication service provided by a certain provider (different from Service IV). The log data was for 1 day in September 2004.

First, we assumed that a link between a pair exists only when there was a communications record for the pair in the one-day log data, and we developed a graph expressing the communications relations between user IDs. In order to eliminate calls that did not arise from social networks, such as calls promoting certain products, a pair was considered to be personal communication only when calls were originated by both parties, each calling the other at least once. From the graph so developed, nodes were sorted according to the number of degree to generate subgraphs. The subgraphs are generated by selecting nodes in accordance with the sorted sequential order of their degree and it becomes the subgraph induced by the selected nodes. We then examined the relationship between the number of links in the induced subgraphs and the total number of calls on links. The result is shown in the left chart in Fig. 2.12. The right chart in Fig. 2.12 shows the results of the induced subgraphs generated by randomly selecting the nodes.

Both results show that the number of calls is proportional to the number of links. Although, in general, the number of calls per link varies greatly from link to link, these results indicate that such a variation does not affect our hypothesis. In other words, the effect of the average values is dominant. These results indirectly verify that the volume of traffic is proportional to the number of links in social networks.

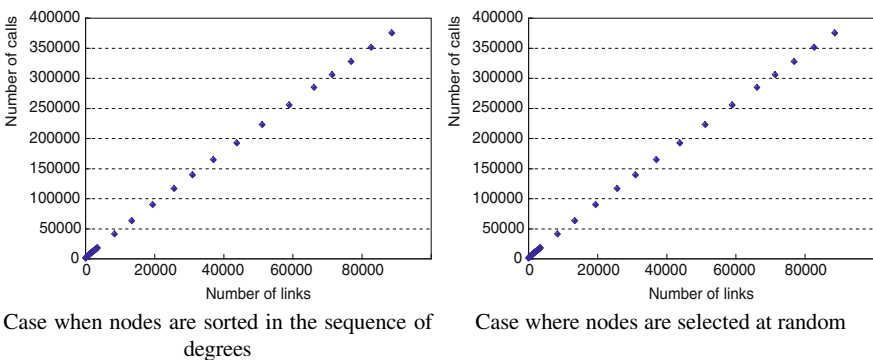


Fig. 2.12 Relationship between the actual number of calls and the number of links in induced subgraphs

B Behavior of $1 - c_x(m)$

Let us examine the behavior of $1 - c_x(m)$ by defining a specific function for $c_x(m)$. If we choose the simplest form that satisfies $c_x(m) \propto m^{1-\delta}$, $c_x(1) = 0$, and $c_x(N) = 1$, then

$$c_x(m) = \left(\frac{m - 1}{N - 1} \right)^{1-\delta}. \tag{2.31}$$

For example, Fig. 2.13 shows the behavior of $c_x(m)$ and $1 - c_x(m)$ for different values of m for the case where the number of potential users, N , is 60,000,000, for δ is set as 0.0 and 0.5, respectively.

Since the range of the number of subscribers, m for Service III (and n for Service I), being considered in this paper is, at most, in the order of several million, we can confirm that the equation, $(1 - c_x(m)) \simeq \text{constant}$, holds on a log scale.

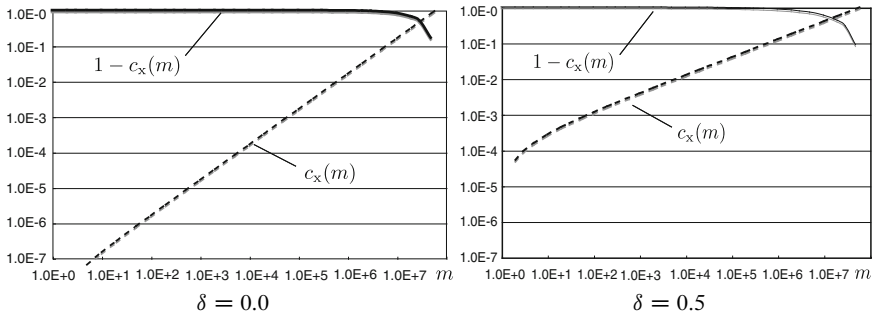


Fig. 2.13 Examples of behavior of $1 - c_x(m)$.