Characteristics of Weighted Email Communications

Yihjia Tsai, Cheng-Chin Lin and Ching-Chang Lin
Department of Computer Science and Information Engineering
Tamkang University
Taiwan, R.O.C.
eplusplus@gmail.com, kevinlin@ieee.org, cclin@mail.tku.edu.tw

***Abstract*— In recent years, network models have been extensively studied. Most of those models are based on the binary status of the existence of a communication link, the intensity of the communication pattern for a given pair of nodes is not taken into consideration. In this paper, we study the statistics of commonly used Email communication and take into account communication frequencies. The resulting weighted Email communication network showed a rather different statistical behavior than the unweighted one. We compare the link and node weight distribution under different assumptions. Finally, weighted clustering coefficient and group growth are introduced and analyzed in the network.**

***Keywords*—weighted networks, Email communications, group growth**

1. Introduction

Network models are commonly used in many branches of science to facilitate the analysis, such as social networks in sociology [1][2], metabolic network [3] and predator-prey network [4] in biology, web page graph [5][6] and Internet [7][8] in computer science, etc. One of the major concerns in network models is that whether it captures some key properties of the subject under study. For example, small-world [9] and scale-free[10] properties.

The concept of small-worlds was originally used in describing human social interconnections. Watts and Strogatz [9] formally proposed two characteristics of small-world networks, that is, network path length and clustering coefficient. Network path length is the average shortest distance between any two nodes in the network. The clustering coefficient C_i of node i is defined as the ratio of actual number of edges connected those nodes between its k neighbors to the number of edges in a fully connected network of k nodes,

$$C(i) = \frac{2E_i}{k(k-1)} \tag{1}$$

where E_i represents the number of edges actually exist between node i's k neighbor. Clustering coefficient of a network is defined as average C_i for all nodes.

A network with a power law distributed degree distribution is called sale-free network. The probability of a node having k degrees is denoted by $P(k)$ and the following is used to describe the power-law.

$$P(k) \sim k^{-\gamma} \tag{2}$$

where γ is the tail index of the power-law distribution. Most of the researches on network modeling have considered a binary network model, that is, the presence or absence of a network link is of major concern. This effectively assumes the weight of all links to be equal to one.

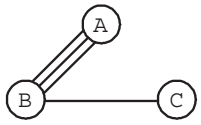


Fig. 1. A network with multi edges between a given pair of nodes.

In fact, many of empirical networks [11][12][13] exhibit the property of intensity. If the links differ in strength, one may indicate this strength by assigning a value to each edge. If we defined the weight as the aggregated value between nodes. As Figure 1 shows that three are three links connecting between node A and B and one link connecting between node B and C. If we defined the weight as number of links between nodes. Figure 1 gives that the weight of $E_{A,B}$ is 3 and is different from $E_{B,C}$. The weighted adjacency matrix W of Figure 1 can then be written as

$$W = \begin{pmatrix} 0 & 3 & 0 \\ 3 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{3}$$

In recent years, empirical network analysis has been widely studied. For example, the world airport networks [11][12][13] represents another case where the weight of links are distinct. The links [11][12] in airport network between two airports represent the capacity of passenger. In [13], the weight defined as number of flights per week between two airports.

We investigate the Email Communication Network (ECN) of TKU. The mail logs from the Email server are studied. The binary network of Email communication has been discussed in [14]. We studied the statistical properties on weighted ECN, which represent social relation between the students. Our analysis shows additional complex properties such as weighted link and node distributions, clustering properties and Email group growth.

2. The Weight

2.1. Weight Definition

A network graph is a connected graph and consists of two sets: nodes and links(edges). Each link is connected by two nodes. An unweighted network graph uses an edge to represent

connections between two nodes. Thus, several links connecting two nodes i and j are seen as one edge $E_{i,j}$. It is commonly used to represent a network by an adjacency matrix $A = [a_{i,j}]$. An element $a_{i,j}$ of the adjacency matrix is either 1 or 0 depending on whether the link between i and j exists or not.

In this paper, we characterize the properties of weighted networks whose links are *weighted*. Each link will be given a weight in network graph. Similar to the adjacency matrix, we define a weighted adjacency matrix $W = [w_{i,j}]$. An element $w_{i,j}$ indicates the weight on the link connecting the node i and j. If $w_{i,j}$ is 0, there is no link between node i and j. A weighted matrix is symmetric, that is $w_{i,j} = w_{j,i}$ and $w_{i,i} = 0$.

2.2. Link Weight Distribution

Most network models have focused on the distribution of node degree, and networks with a power law degree distribution are said to have the scale-free property [10]. That is, when we denote the probability of a node has k_i links as $P(k_i)$, we can use $P(k_i) \sim k_i^{-\gamma}$ to describe the power-law distribution. The tail index γ is of major concern in those networks. Similar to degree distribution, we are also interested in the link weight distribution in a weighted network. Define $P(w_i)$ as the probability of a link with weight w_i, the link weight distribution can be studied in addition to node degree distribution.

2.3. Node Weight

A node i obtains its weight information from all weighted links between its neighbors. In unweighted networks, nodes with large number of connectivities are of major focus, but in weighted networks, nodes with the same degree of connectivity may not have the same weights, In [11], a node weight Q_i is defined to be the sum of all w_j form all out-going links of node i.

$$Q_i = \sum_j a_{i,j} w_{i,j} \tag{4}$$

We propose a node weight q_i defined as

$$q_i = \frac{\sum_{j \in V_i} w_{i,j}}{|V_i|} \tag{5}$$

where V_i is the set neighbor nodes of i, and $|V_i|$ represents the number of elements in the set. q_i is in some sense a normalized Q_i and it provides a good measure of node weights on a per link basis.

2.4. The Weighted Clustering Coefficient

The clustering coefficient $C_w(i)$ of weighted networks presents strengths between node i's neighbors. In $C(i)$, a link only indicate a connectivity between two nodes and no strength shows between nodes. But, in weighted network, each link with a weight shows the strength between two nodes. Therefore, if $C_w(i) > C_w(j)$ means that node i has stronger relationship between neighbors than node j even $C(i) = C(j)$, because $C(i)$, as Equation (1), only considered number of links. Barret et al.[11] proposed a measure of the clustering that combines the topology information with weight distribution of the network. The clustering coefficient defined as

$$C^w(i) = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{w_{ij} + w_{ih}}{2} a_{ij} a_{ih} a_{jh} \tag{6}$$

It counts the two node participating links of the node i for each triple formed in the neighborhood of the node i.

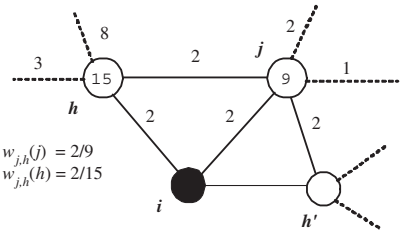


Fig. 2. The subgraph of node i. Sum of link weights of nodes j and h are 9 and 15, respectively. The $w_{j,h}$ shows different strengths to nodes j and h, where $w_{j,h}(j) > w_{j,h}(h)$.

The definition of clustering coefficient from Watts [9] gives a measure of tightness between a group of nodes in the network. Similarly, we proposed a weighted clustering coefficient C_w in term of weighted links. A link with weight $w_{j,h}$ can mean differently to node j and h when taken into consideration the node weights. Therefor, we define a weight ratio $w_{j,h}(j)$ for node j to be

$$w_{j,h}(j) = \frac{w_{j,h}}{Q_j} \tag{7}$$

Figure 2 demonstrates an example of different weight ratios of a common edge viewing from its end nodes. The edge $E_{j,h}$ has a weight of $w_{j,h}$=2 and node h has a node weight Q_h=15 while the node weight of j is 9. Therefore, viewing from the position of node h, the edge $E_{j,h}$ is only $w_{j,h}(h) = w_{j,h}/Q_h = 2/15$ of its node weights. Similar calculation results in the weight ratio of $E_{j,h}$ from node j as $w_{j,h}(j) = 2/9$.

After we known the relation of link strength to node weight. Similar to Equation (1), we calculate all links that connected between node i's neighbors, such as $E_{j,h}$ and $E_{j,h'}$. The $C_w(i)$ can be defined as the ratio of all link strengths between node i's neighbors to number of neighbors of node i. In other words is the probability of the strengths of your friends between your friends. Thus the $C_w(i)$ can be denoted as

$$C_w(i) = \frac{\sum_{j \in V_i} \sum_{h \in V_i} w_{j,h}(j)}{|V_i|} \tag{8}$$

In order to prevent $C_w(i) \geq 1$, the $C_w(i)$ is normalized by number of neighbor. It ensures that $C_w(i) \in (0, 1)$. That is, the $C_w(i)$ can be compared between nodes. The network weighted clustering coefficient C_w is also an average $C_w(i)$ for all i in the network. And, C_w also can be compared between weighted networks.

3. Empirical Results

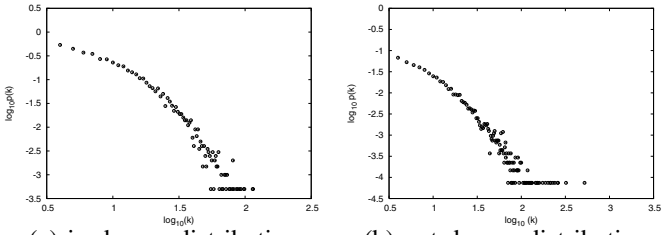
We investigated the empirical data from the TKU's Email server. The data was collected from Sep. 2001 to Oct. 2002.

Only internal mails were selected for constructing the Email communication network (ECN). More than 3,600,000 internal Emails were selected with 13,021 Email accounts. In the ECN, each account represents as a node and a link $E_{i,j}$ means an Email sent between nodes i and j. We define the weight in ECN is number of communications. That is, each given weight $w_{i,j}$ of link $E_{i,j}$ is the number of bi-direction Emails between i and j.

3.1. Degree Distribution

The distributions of in-out degree are studied and shown in Figure 3. Direction is one of important characteristics in Email communications. An in-link of node i represents an Email send to node i and an out-link of node i is an Email send from node i. The in-degree is defined as number of nodes that ever sent Email to node i and out-degree is number of nodes that had been received Email from node i.

The degree distribution of Email communication is discussed in [14]. The scale-free network with a tail index by a power law can be explained a growing network with preferential attachment. The tail index γ of TKU's Email communication network is 2.01[14]. According to the directional property of Emails, we shows the in-out degree distributions in Figure 3. We find that the degree distributions of both directions follows a power law. Qualitatively the power-law properties are presented while the tail indexes are different. The γ of in-degree distribution is 2.48 and the γ of out-degree distribution is 1.72.



(a) in-degree distribution (b) out-degree distribution

Fig. 3. The log-log plot of degree distributions of ECN. k is number of degree and $P(k)$ represents of probability of k. (a) The degree distribution of in-degree. (b) The degree distribution of out-degree.

3.2. Link Weight Distribution

In a weighted network, each link has a weight representing its communication strength. Figure 4 shows the weight distribution of the weighted ECN. The distribution is characterized by a tail index of γ=0.73. For weighted ECN, the general trend is that nodes with large degree have large average weights, however, this trend does not hold true when node degree exceeds a certain threshold. As illustrated in Figure 5, for the ECN under study, 70 is the threshold.

3.3. Node Weight Distribution

We analyzed the ECN by two equations of node weight, Q_i and q_i. Figure 6 shows the distributions of nodes weight of ECN. q_i is node weight and $P(q_i)$ represents of probability of

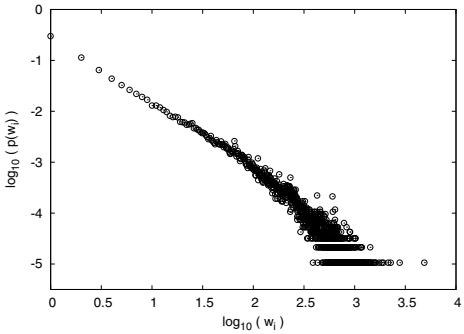


Fig. 4. The weighted link distributions of ECN. w_i is link of weight and $P(q_i)$ represents of probability of q_i.

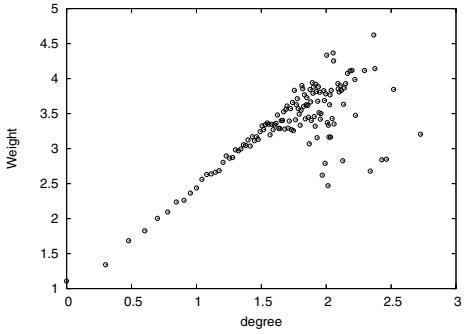
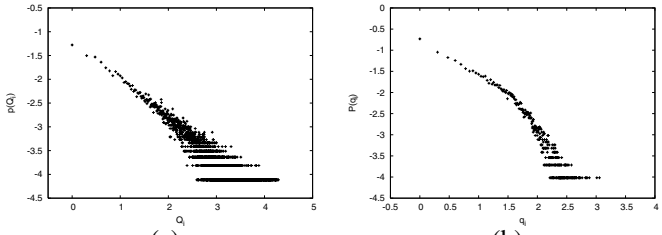


Fig. 5. The Log-log plot of the relation between degree k_i and average weight of degree k_i.

q_i. In Figure 6(a) shows a power-law distribution of the node weight by Equation (4). It not easy to find out the properties of node weight. As section 2.3 mentioned, the node weight should be considered with node degree. Figure 6(b) shows the node weight distribution with Equation (5), it explicit shows that two lines with different γ's in the log-log plot. Before $q_i < 1.5$ the γ is 0.84 and γ=1.85 where $q_i \geq 1.5$.



(a) (b)

Fig. 6. The log-log plot of distributions of nodes weight of ECN. q_i is node weight and $P(q_i)$ represents of probability of q_i. (a) The node weight distribution by sum of weight of all links of node i. (b)The node weight distribution by Equation (5). Two lines with different γ show in this plot. Before $q_i < 1.5$ the γ is 0.84 and γ=1.85 where $q_i \geq 1.5$.

3.4. The Weighted Clustering Coefficient C_w

From social network analysis, an important property of the studied networks is the clustering coefficient. The clustering coefficient,$C_w(i)$, captures the probability of local relationship

of a node. The average clustering coefficient, C_w, presents the global density of all nodes in the network. The characteristics of clustering of unweighted Email communications are discussed in [14]. In Table I shows clustering coefficient of weighted/unweighted ECN. According to Equation (8), the weighted clustering coefficient C_w of the ECN is 0.3008. The

TABLE I
CLUSTERING COEFFICIENT OF THE ECN

	unweighted ECN	weighted ECN
C [14]	0.24	-
C_w	-	0.3008

other considerations which are used to analysis the clustering coefficient is the distribution of $C_w(i)$ and the correlation between $C_w(i)$ and degree. First we study the distribution of $C_w(i)$ as shown in Figure 7. The $C_w(i)$ decreasing very fast while $C_w(i)$ exceeds a threshold. Before the threshold, $P(C_w)$ only a small varying with C_w.

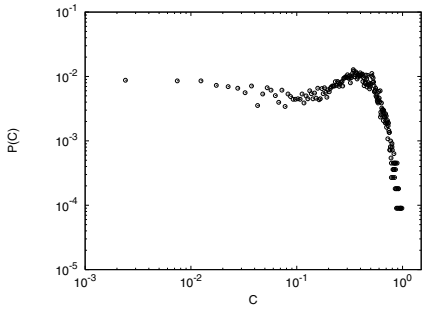


Fig. 7. The correlation between $C_w(i)$ and probability of $C_w(i)$.

Figure 8 shows the correlation between average node weight of the nodes with degree k and node degree k. As illustrated in Figure 8, the node weights increasing with node degree increasing until degree exceeds 70. From Figure 8 and Equation (8), when degree increasing, the neighbors of node i increasing strength on node i's neighbors.

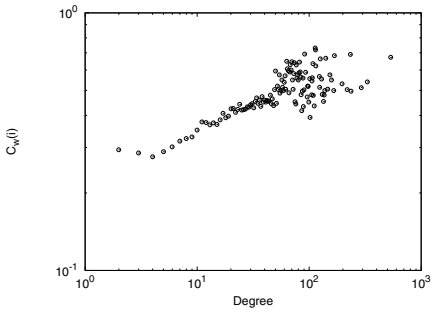
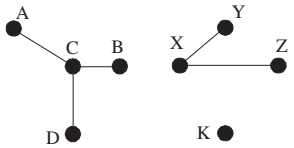


Fig. 8. The correlation between average $C_w(i)$ and node degree k_i.

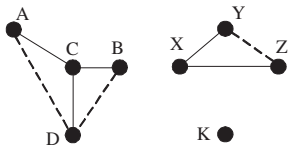
4. EMAIL GROUP GROWTH

4.1. Group

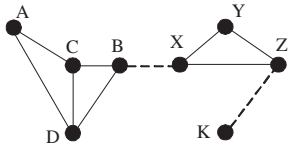
A group contains of nodes and links. The nodes will make up some groups according to the connection of links. Therefore, in the same group, any node can reach another node along links. For example, in Figure 9(a), node A, B, C and D are in same group. Node K and X belong to different groups.



(a) Three groups in the network



(b) Intra-group communications



(c) Inter-group communications

Fig. 9. Group communications. (a) Three group, {A,B,C,D}, {X,Y,Z} and {K}, in the network. (b)(c) Intra-group and Inter-group communications.

4.2. Group Communications

The group communication in our paper defined as two types, inter- and intra- group communications. In Email networks, Emails communicate between nodes. Thus, the Emails/links are represented the communications between nodes and groups. The link $E_{i,j}$ which nodes i and j belong to the same group is defined as the intra-group communication. For example, in Figure 9(a) shows three groups, {A,B,C,D}, {X,Y,Z} and {K}. Figure 9(b) shows that $E_{A,D}$, $E_{B,D}$ and $E_{Y,Z}$ are the intra-group communications. The links which connect between groups are the inter-group communication, such as $E_{B,X}$ and $E_{K,Z}$ are inter-group communication shows in Figure 9(c) .

Behavior of the group communications influences the characteristics of Email group growth. The speed of Email group growth relates to the types of the group communication. During a time period, the group growth is very slow, if most of the Emails are all conveyed among members of group. That is, the most of the Emails are intra-group communication. Certainly, most of Emails are inter-group communication will accelerate the speed of group combined.

5. THE PATTERN OF EMAIL GROUP GROWTH

The link set in the ECN can be represented by two kinds of types, weighted link set and unweighted link set. In the ECN, we have a node set including 13021 nodes and a link set. The link set can be represented to weighted and unweighted links. For example, the links in Figure 1 can be represented as four unweighted links or two weighted links. Thus the link set in

the ECN can be presented by two types, one is *weighted link set* including 94041 links and the other is *unweighted link set* including more than 3,600,000 links.

We observe the group variation by adding links into the network. The nodes which are connected or an isolated node is called a *group* in network graphs. A group consists of at least one node. Therefore, the network graph will begin with 13021 groups. With the joining of links, group mutually combines with group. Finally, the 13021 nodes becomes to a connected network graph. According to the way of link joining, there should be different group variations. Two patterns are considered, there are link weight increasing, link weight decreasing. The scenario follows the steps as bellow

1) The network graph begins with 13021 isolated nodes.
2) select a link from link set.
3) Add the link into the network, and then recompute amount of groups.
4) repeat step 2-3 until the network is connected.

First, we observe group variation with weighted link. First, the weighted link set is sorted by weight increasing. And second is the weighted link set sorted by weight decreasing. Following the scenario steps, the Figure 10 shows group variations of two patterns. The group variation in Figure 10

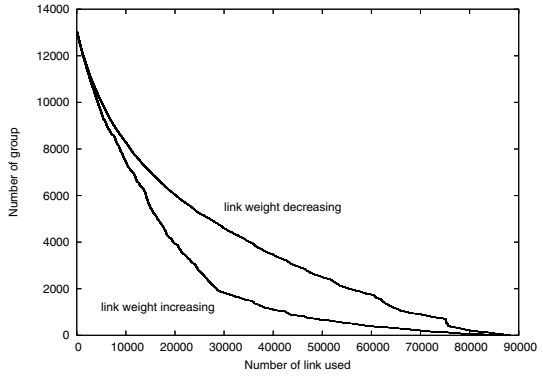


Fig. 10. The group variation with link adding. The links in weighted link set are sorted by link weight increasing or decreasing. The x-axis is number of link used. The y-axis is number of group.

shows two curves as link weight increasing and link weight decreasing respectively. Two curves starts with 13021 groups and than groups are combined with link adding. Finally there is only one group in the network, that is, the network is a connected network. In Figure 10, the curve of link weight increasing is decreasing fast than link weight decreasing. That is, the links with weight increasing create a large number of groups in the network at beginning time. However, the links with weight decreasing do not combine groups very fast. Thus the heavy weight of links have high probability connecting nodes in the same group.

According to Figure 10, the topology of the ECN is clusters topology. If the ECN is an intensive or sparse network, the β of two curves should be very close. In fact, in Figure 10, two curves differ by 2500 groups after 30,000 links has already been added. Therefore, the ECN topology is clusters.

6. Conclusion

In this paper, we present characteristics of the Email communication network with different types of weight and group growth. The metrics of weight allows to characterize the statistical properties of the strength of links and nodes. We analyzed the ECN network quantities as link/node weight distribution, the correlation between degree and different weights. We also calculated the clustering coefficients of nodes and the ECN network. These results give us clues to understanding the role of weight in the network topology.

The node weight, q_i, and the clustering coefficient, C_w, we proposed are use to investigate the ECN network. The proposed node weight shows the relationship between weight and network structure. Two γ values exhibited different weight distributions. We focus on the proportion of weight is communicated with our common friends. The calculated results approach to our idea.

The group growth exhibits the Email network topology. Two types of group growth give the clues of network topology. By the high clustering coefficient[14] and group growth, the topology of the ECN is a cluster network.

The weighted communications and the group growth can be used for investigate the Email applications. For example, the computer viruses disseminate on Internet via Email communications. The results of our paper can be used for the Email researches such as Email viruses, human communications on Internet. One of our future works is the contagion of computer viruses through Email communications.

References

[1] M. E. J. Newman, S. Forrest and J. Balthrop, "Email networks and the spread of computer virues," Phys. Rev. E 66, 035101, 2002.
[2] D. J. Watts, P. S. Dodds and M. E. J. Newman, "Identity and search in social networks," Science 296, pp.1302-1305, 2002.
[3] H. Jeong, B. Tombor, R. Alert, Z. N. Oltvai and A.-L. Barabá*i*, "The large-scale organizaton of metabolic networks," Nature 407, pp.651-654, 2000
[4] S. L. Pimm, "Food Webs", University of Chicago Press, Chicago, 2nd ed. 2002.
[5] M. E. Crovella, M. S. Taqqu, A. Bestavros, "Heavy-tailed probability distributions in the World Wide Web," in A Practical Guide To Heavy Tails, chapter 1, pp.3-25, Chapman & Hall, New York, 1998
[6] R. Albert, H. Jeong, A.-L. Barabási, "Internet: Diameter of the World Wide Web," Nature 401, pp.130-131, 1999
[7] T. Bu and D. Towsley, "On distinguishing between internet power law topology generators," IEEE INFOCOM, USA IEEE Computer Society Press, Los Alamitos, CA, USA, 2002.
[8] K. L. Calvert, M.B. Doar, and E. W. Zegura, "Modeling Internet topology," IEEE Communicaton Magazine, 35(6): 160-163, 1997.
[9] D. J. Watts, D. H. Strogatz, "Collective dyanmics of 'small-world' networks," Nature 393, pp.440-442, 1998.
[10] R. Albert, A.-L. Barabá*i*, "Statistical mechanies of complex networks," Reviews of Modern Physics, Volumne 74, pp.48-97, Jan. 2002
[11] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, Proc. Natl. Acad. Sci. (USA) 101 3747 , 2004
[12] S. R. Guimera, M. Sales-Pardo, and L. A. N. Amaral. "Structure and Efficiency of the World-Wide Airport Network," cond-mat/0312535, Dec. 2003
[13] G. Bagler, "Analysis of the Airport Network of India as a complex weighted netwrok," arXiv:cond-mat/0409773, 2004
[14] Y. Tsai, C.-C. Lin, P.-N. Hsiao "Modeling Email Communications," IEICE Transactions on Information and Systems Vol.E87-D No.6 p.1438, 2004