Network Analysis Summary

Lei Meng

1 Motivation

In our analysis, we aim to answering the following two questions.

- Given a type of communication, whether networks constructed from close time are more similar or networks constructed from far apart time are more similar.
- Whether networks formed from different communication type are similar or different. How much similar do they have, if they are not completely different?

2 Method

We collect five types of communication from cellphone users as shown in Table 1. The data in our analysis is gathered from September 2011 to Feb 2013.

1	SMS	
2	Phone Call	
3	Email	Not including group emails
4	Facebook	Comments, tags, public messages
5	Bluetooth	At least one person detects the other with RSSI greater than some threshold

Table 1: Communication types

2.1 Network Construction

- **Network**: We divide the whole study period in T non-overlapping intervals of length Δt and construct a network $G_t^x = (V, E)$ for each communication type $x \in X$ in each interval $(t, t + \Delta t)$, $t \in \{t_0, t_0 + \Delta t, t_0 + 2\Delta t, ...\}$. Since there is no directionality in proximity, we simply treat all the networks as undirected. $w_t^x(i, j)$ is the number of communications occurred between i and j in network G_t^x. If $w_t^x(i, j)$ is greater than some threshold w, $e(i, j) \in E$, otherwise $e(i, j) \notin E$
- **Node**: Each user in our study represents a node in the network. Users are not included if they dropped out of the study in the middle so that all the networks contain fixed nodes.
- **Bluetooth Network** : If two users i, j detect each other continuously for at least 6 minutes satisfying some RSSI threshold t (at least once in the first 3 minutes and at least once in the second 3 minutes), we consider i and j has a bluetooth communication from the time they detect each other to the time they do not detect each other. This method can eliminate the situation that two people are simply walking by each other without actually spending some time together.

2.2 Network Properites and Similarity Measurements

2.2.1 Network Properties

For each constructed network, we calculate the following properties.

- Number of nodes $|V|$ in the network.
- Number of edges $|E|$ in the network.
- Number of nodes in the largest connected component $|V_{LCC}|$.
- Number of edges in the largest connected component $|E_{LCC}|$.

- Percentage of nodes in the network that are also in the largest connected component $\frac{|V_{LCC}|}{|V|}$.
- The density of the network $\frac{2|E|}{|V|(|V|-1)}$.

2.2.2 Similarity Measurement

For the purpose of comparing the similarity of two networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, we use the following measurement.

- Edge intersection (I), the number of shared edges of the two networks $|E_1 \cap E_2|$.
- Relative edge intersection (RI), the number of shared edges divided by number of unique edges in both networks $\frac{|E_1 \cap E_2|}{|E_1 \cup E_2|}$.
- GDD-agreement (GDD)
- Pearson correlation of degree distribution (PDD)

2.2.3 Parameters

Table 2 and Table 3 show the different parameters Δt, w and r studied in our analysis. Δt is the number of days the networks are constructed from, w is the minimum number of communication occurred in the given time interval between two nodes i, j in order to form an edge in the network and r is the minimum signal strength of Bluetooth detected by either one of the person to ensure that the two person are within a certain distance.

Time interval(Δt)	Threshold(w)
One week	1
One month	1
One month	2
One month	3
One month	4
Three months	1
Three months	2
Three months	3
Three months	6
Three months	9
Three months	12
Six months	1
Six months	3
Six months	6
Six months	12
Six months	18
Six months	24

Table 2: Thresholds studied in our analysis.Δt is the number of consecutive days used to construct the network. w is the minimum communication occurred to connect two users in the network.

RSSI threshold(r)
-50
-55
-65

Table 3: Threshold r is used to evaluate the distance between two users.

2.3 Network Comparison

To answer the two questions we mentioned previously, we conduct two kinds of network comparison, homogenous network comparison and heterogenous network comparison. Network comparison occurred only on networks of the same parameters. Homogenous network comparison is to compare networks of the same type but from different time to answer whether networks is more similar to another of the same type around the time or not. For each communication type and given parameters, we compare each network from different time with every network based on the similarity measure we discussed previously. We construct a matrix for each similarity measure, the columns are networks in the order of increasing time. The rows are the same as the columns. The diagonal of each matrix is 1 since each network is the most similar to itself. An element e_{ij} in the matrix would be the similarity result of network i and network j. Heterogenous network comparison is to compare networks of different type but from the same time. By conducting this comparison, we would be able to find if one type of communication is more similar to another type.

3 Result

3.1 Network Properites

Figures Sup.1-3 show the network density over time for different network types and different parameters. We treat Bluetooth networks of different RSSI threshold r as different type of networks. Thus, we have seven types of networks to study. Figures Sup.4-6 show the percentage of nodes from the network in the largest connected component over time for different network types and different parameters..

- From the density graphs, we can see that the density of all the networks is below 0.2 which indicating our networks are sparse.
- Network constructs from six months are able to show the effect of summer break(three months) but is not able to show what happens during a shorter period of time such as winter break (one month). Network constructs from three months is able to show summer and winter breaks but not able to show spring break(one week).Thus, we chose Δt=1 week and Δt=1 month. Table 1 shows the holidays and their durations and the time each holiday occurred for different time interval Δt.
- For Bluetooth networks, as r increases the size of the largest connected component increases as well for the same parameters. When $r = -50$, the size of the connected component is very small compared to -55 and -65. Thus, we only focus on networks with $r = -55$ and $r = -65$ for the rest of study.
- For network with Δt=1 month, as w increases, the size of the largest connected component decreases for the same network type. When $w = 4$, the size of the largest connected component especially for Facebook network is very small. Thus we will focus on $w \leq 3$ for $\Delta t = 1$ month.
- For networks with Δt=1 month, the longer the holidays are, the smaller the size of the largest connected component is. For example, the size of the largest connected component in the summer break is smaller than the winter break as shown in Figure 4 Row 1 (SMS) and Column (b).
- For networks with Δt=1 week, we can observe not only summer break and winter break, but also shorter breaks such as Spring break. For example, in Figure 4 Row 4 (Facebook) and Column (a), Spring break has a smaller fraction of nodes in the largest connected component than the time before and after.
- In the summer, the size of the largest connected component of different network types except Facebook network are very small indicating that people do not communicate much through text message, phone call, in-person meeting or email when they are at long holidays. However, they communicate through Facebook as shown in Figure 4 Column (b).
- For Facebook network as shown in Figure Sup.4 Row 4 (Facebook) and Columns(b, c, d, e), as threshold w increases, the size of the largest connected component drops quickly which indicating that communication through Facebook does not occur often. For other types of network, the increase of threshold w does not impact the size of the largest connected component that much as it does on Facebook network.
- For Email network, the size of largest connected component is very small for any threshold indicating that most of the nodes in this network are isolated. In another word, most people do not communicate through email.

In summary, we use Δt=1 week, $w = 1$ and Δt=1 month, $w \leq 3$ in the following analysis. We define Bluetooth (-55) as Bluetooth network with $r = -55$ and Bluetooth (-65) as Bluetooth network with $r = -65$. For simplicity, we treat Bluetooth (-55) and Bluetooth (-65) as different type of communication networks.

3.2 Homogeneous Network

In this section, we compare networks of the same type from different time.

3.2.1 Intersection & Relative Intersection

The first method we use to examine the similarity of two networks is the number of shared edges between two networks. Figure Sup.7 shows the intersection matrix for different types of networks with different parameters. For example, the matrix in Row 1 (SMS) and Column (a) is generated by computing the number of shared edges of every pair of SMS networks from different time. We also look at the the intersection relative to the size of the two networks. Figure Sup.8 shows matrices generated from calculating the relative intersection of each pair of networks from the same type but different time.

For intersection matrices shown in Figure 7, for $\Delta t = 1$ month, as threshold w increases, we can see the colors of the graphs get lighter for the same type of network, indicating that the intersections get smaller. For relative intersection matrices shown in Figure 8, as w increases, the differences of the same type of network are not significant. Thus, we choose $w = 3$ for $\Delta t = 1$ month in the following analysis along with $\Delta t = 1$ week, $w = 1$. In the following sections, we refer network with Δt=1 week, $w = 1$ as weekly network and network with Δt=1 month, $w \leq 3$ as monthly network.

- In the intersection matrices, the maximum value estimated from the legend shows the maximum size of the network. For example, for networks ($\Delta t = 1$ week, $w = 1$) as shown in Figure 7 Column (a), the maximum size of Facebook network is 820 and Bluetooth(-65) is 710. SMS, Phone Call and Email network is much smaller than Facebook and Bluetooth networks. The ordering of the maximum sizes of each type of network is Facebook > Bluetooth (-65) > Bluetooth (-55) > SMS > Phone Call > Email for weekly networks. For networks ($\Delta t = 1$ month, $w = 3$) as shown in Figure 7 column (d), the ordering of the maximum sizes of networks of different types is the same as the one from $\Delta t = 1$ week, $w = 1$ for SMS, Phone Call, Email, Bluetooth (-55) and Bluetooth (-65). However, the maximum Facebook network size is much smaller indicating that the increasing threshold w prune lots of edges that formed in the weekly graphs. The ordering of the maximum size of each type of network is Bluetooth(-65) > Bluetooth(-55) > SMS > Facebook > Phone Call > Email for the selected parameters.

- The intersection graphs in Figure Sup.7 show that the color is darker around the diagonal and the further away from the diagonal, the lighter the color for all networks except Facebook network. This founding indicates that networks close in time are more similar, and are more different if far in time for SMS, Phone Call, Email, Bluetooth networks.

- The Facebook network ($\Delta t = 1$ week, $w = 1$), the intersection graph shown in Figure 7 Row 4(facebook) and Column (a) has the darkest spot at (23, 23). Further looking at Figure 11 (a), the size of weekly Facebook network starts to increase at week 21 and reaches its peak at week 23. Week 21 is one week after the second semester starts. There is a second peak between week 50 and 55 which are the weeks after the summer break.

- As shown in Figure 7 Row 5,6 (Bluetooth (-55) and Bluetooth (-65)) and Column(a), the color is darker along the diagonal from week 5 to week 36 (before summer break) than other time. This might students are not in the same class or dorm after the first year, thus they do not meet as much as before.

- In Figure 7 Row 4(Facebook) and Column (d), networks are more similar between month 0-5 and month 12-17. Month 0-5 are the initial stage and month 12-17 are the periods after summer break.

- Figure 8 Row 4 (Facebook) and Column (c) shows that Facebook networks have few overlap relative to the size of the network whether time is close or not. The maximum overlap is 20.5% (excluding diagonal values).

- Summer looks more similar since the size of the network is small.

- For networks ($\Delta t = 1$ week, $w = 1$), the relative edge intersection graphs as shown in Figure 8 Column(a), the relative overlap between a network in break time and a network not in break time is nontrivial in SMS and Phone Call but almost zero in bluetooth network. For example, the relative overlap between week 10 and 45 is 10.8% in SMS and 10.7% in Phone Call, but 0.6% and 0.1% in Bluetooth (-55) and Bluetooth(-65) respectively. This indicates that some portion of people still communicate consistently over both the holidays and working days. But they tend not to meet over the break.

- For Facebook network ($\Delta t = 1$ month, $w = 3$), networks are more similar among month 0-5 in terms of edge intersection in Figure Sup.7 but not in terms of relative edge intersection in Figure Sup.8. Networks among week 6-17 is more similar in terms of relative edge intersection but not in terms of absolute edge intersection. The reason for this is the size of the network in the first five months is almost twice the size of the networks after week 5 as shown in Figure Sup.11 (b).

3.2.2 GDD-agreement

Figure 9 shows the results of comparing homogenous networks based on GDD-agreement for $\Delta t = 1$ week, $w = 1$ and $\Delta t = 1$ month, $w = 3$.

- In Column (a), Bluetooth(-65) networks are similar if their time is close except breaks. Bluetooth(-55) networks are also similar if their time is close except breaks. However, the time span of such similarity is smaller than network Bluetooth (-65). SMS and Phone Call are more similar after week 36(start of summer break). Friendship transitivity might has formed and stabled.
- In Column (b), Facebook networks are more similar after month 5.

3.2.3 Pearson correlation of degree distribution

Figure Sup.10 shows the results of comparing homogenous networks based on Pearson correlation of degree distribution for different parameters. $\Delta t = 1$ week, $w = 1$ and $\Delta t = 1$ month, $w = 3$.

- Column (a) presents the results from $\Delta t = 1$ week, $w = 1$. Networks from week 16-18, 37-49, 68-70 is not considered in the comparison of Phone Call networks, Bluetooth(-55) networks and Bluetooth (-65) networks, since they are holidays and the network sizes are very small(< 15). Email network is excluded in the results since the size of its network is also very small.
- Column (b) presents the results from $\Delta t = 1$ month, $w = 3$. Networks from month 10 in Phone Call network and month 9-10 in Bluetooth(-55) network and Bluetooth(-65) network is not considered in the comparison for the same reason as mentioned above. Email is also excluded in the results for the same reason.
- Networks that are close in time are also similar.
- The reason that sometimes the Pearson correlation of Bluetooth (-65) network and other types of network network is -1 is because SMS, Phone Call, Email, Facebook, Bluetooth(-55) follow a power law degree distribution but Bluetooth (-65) sometimes has skewed normal degree distribution.

We compared networks of the same types using four different similarity measures. In order to understand if different similarity measures are correlated, in another word, if one measure indicates strong similarity of two networks, does the others indicate the same strong similarity or the opposite, we calculate the Pearson correlation of each pair of the similarity results from the four measures. For a particular type of network and selected parameters, each similarity measure f generate an symmetric matrix $M_{T\times T}$. For $e_{ij} \in M_{T\times T}$ $(0 \leq i \leq T-1,\ 0 \leq j \leq T-1)$, e_{ij} is the value of comparing network from time i and network from time j based on measure f. We took the values of upper triangle (excluding the diagonal) in the matrix and form a vector $(e_{01}, ..., e_{0(T-1)}, ..., e_{i(i+1)}, ..., e_{i(T-1)}, ..., e_{(T-2)(T-1)})$. And then we calculate the Pearson correlation of the two vectors, each from a different measure.

Table Sup.2 shows the correlation of pairwise similarity measure in comparing networks of the same type with Δt=1 week, $w = 1$. Email network is excluded since it has been excluded in the comparison of Pearson correlation of degree distribution. For example, the Pearson correlation of RI and I on SMS networks is generated by computing the Pearson correlation of Figure 7 Row 1 (SMS) Column (a) and Figure 8 Row1 (SMS) Column (a). From the results, we can see that Intersection and Relative Intersection are strongly correlated in all different types of networks. Pearson correlation and GDD-agreement are strongly correlated in all network types except SMS. In Bluetooth(-55) networks, Intersection is correlated with PDD and GDD. In Bluetooth(-65) networks, Intersection is correlated with GDD, RI is correlated with GDD and PDD.

Table Sup.3 shows the correlation of pairwise similarity measure in comparing networks of the same type with $\Delta t = 1$ month, $w = 3$. For Bluetooth (-65) network, all four similarity measure are correlated with each other. For the rest types of network, I is strongly correlated with RI. GDD and PDD are correlated in Facebook, Bluetooth (-55). In Facebook network, RI is correlated with GDD.

3.3 Heterogeneous Network

In order to find if there exists two types of networks more similar than others, we calculate similarity on each pair of the network types. We again apply I, RI, GDD, and PDD four similarity measures. Figure Sup.12 and Figure Sup.13 show the comparison results for weekly and monthly networks respectively. The columns are the time and rows are network pairs to compare. For pearson correlation of degree distribution in Figure Sup.12 (d), week 16-18, 37-49, 68-70 are not computed since they are holidays and the network sizes are very small(< 15). Moreover, Email network is also not compared to any other networks due to its small size. In Figure Sup.13 (d), for pearson correlation of degree distribution, month 9-10 is not calculated and Email network is not considered in the comparison for the same reason as discussed above. Table Sup.4 and Table Sup.5 show the average and standard deviation of each row in each matrix in Figure 12 and Figure 13 respectively. The numbers in bold are the largest average in each column.

- It is not surprising that Bluetooth(-55) and Bluetooth(-65) have a higher intersection than other network pairs. If user i can detect user j with RSSI > -55, i can detect user j with RSSI > -65. Bluetooth(-65) contains Bluetooth(-55).
- Among all the network types, Bluetooth(-55) and Bluetooth(-65) shared the most edges with SMS except itself. This indicates that people who text each other are often co-located.

- Among all the network types, SMS shared the most edges with Phone Call indicating that if two people call each other, they tend to text as well.
- Bluetooth(-65) has a lowest GDD-agreement with other networks types. This might be due to the non power law distribution of Bluetooth(-65).

In order to see if any two similarity measurement are correlated in comparing heterogenous network, we compute the correlation between any two similarity result on heterogenous network comparison. The correlation results are shown in Table 6.

- For weekly networks, I and RI, RI and GDD, GDD and PDD are correlated.
- For monthly networks, all similarity measures are correlated.

3.4 Node degree evolution

In order to understand how each node's degree changes over time, we plots a matrix for each type of communication network as shown in Figure 50 and Figure 51 for weekly and monthly interval respectively. Each column is time and each row is a node. The row index on different figures represents the same nodes.

- For SMS and Phone Call, there are a few nodes that have high degrees over the whole study period while the rest of them have low degrees. Each node's degree at different time is quite similar except for holidays that they do not communicate much.
- For Email, the communication is not consistent, it is hard to find a node that have email communication through the whole study period. Some people do not communication through email at all.
- For Facebook, lots of nodes have a high degree at week 23 in Figure Sup.14 (Facebook) which is consistent with Figure 11 (weekly) where the size of the network is at its peak in week 23. The large degree of a node is quite bursty and is not persistent over the time.

We sort the node in the order of their average degree in SMS networks with the corresponding parameters. And redraw the node degree evolution in the same sorted order. The results are shown in Figure 16 and Figure 17. If a person's has a larger degree in SMS network, the person tends to have larger degree in Phone Call network, Bluetooth (-55) and Bluetooth (-65) network but not necessarily in Facebook and Email network.