Impact of Attributes on Group Formation

Ashwin Bahulkar*, Boleslaw K. Szymanski*, Kevin Chan† and Omar Lizardo‡
*Department of Computer Science and Network Science and Technology Center
Rensselaer Polytechnic Institute, 110 8th Street, Troy NY 12180, USA
Email: bahula@rpi.edu, szymab@rpi.edu
†US Army Research Laboratory,
Adelphi, MD 20783, USA
Email: kevin.s.chan.civ@mail.mil
‡Department of Sociology
University of Notre Dame, Notre Dame, IN 46556
Email: lizardo@nd.edu

***Abstract*—Previous work has shown that selectivity based on opinions and values of attributes is an important tie-formation mechanism in human social networks. Less well-known is how selectivity influences the formation and composition of whole groups in which interactions extend beyond the dyads. To address this question, we use data from the NetSense study consisting of a multi-layer (nomination, communication, co-location) network of university students. We examine how group formation differs from tie-formation in terms of the role of selectivity based on opinions and attributes. In addition, we show how levels of such selectivity varies between groups formed to meet different needs.**

***Index Terms*—Groups, communities, social networks, social influence in groups**

I. Introduction

It is well-known that self-selection based on common values of attributes (homophily) is an influential tie-formation selection mechanism, as noted in [5], [10], and [4]. However, in spite of a perennial concern among social scientists on the nature and evolution of social groups [2], very little work has been done on analogous selection mechanisms at the group level (for some early sociological work on the subject see [9]). Consequently, very little is known about the role common-attribute homophily plays in the formation and evolution of groups. This is an important question given the status of groups as one of the primary forums within which human interaction and sociability take place [3]. Important processes of interest to network scientists such as influence, opinion formation, and opinion-change also happen at the level of groups and not just isolated dyads [17].

In this work, we explore how self-selection based on common attributes impacts the formation of groups, using an innovative method of identifying face-to-face groups using Bluetooth co-location data. We then explore differences in the ways selectivity based on common attributes influences two different network outcomes: frequency of communication and common group membership as revealed by co-location links among study participants. We first hypothesize and then confirm that attribute values based selectivity plays a stronger role in tie-formation based on communication links than in determining common group memberships based on co-location. Finally, we measure how these groups are polarized in terms of multiple related attribute values and opinions.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain

II. Data

The NetSense data used in this work is derived from a multi-semester observations of a subset of one class sample of university students. The data are multi-dimensional. One part consists of survey-based self-reports containing the students' views and opinions on various social political issues, their personal interests and socio-demographic background. These data where collected at the beginning of every school semester from the Fall 2011 to Spring 2013 [16]. In addition, also at the beginning of each semester, students are asked to list their social contacts with whom they interact most frequently.

These self-report data are augmented with non-obtrusive records of electronic communications obtained from smartphones (calls and texts) exchanged between study participants. Finally, Bluetooth interactions indicating spatial proximity between study participants were also recorded. We identify an evolving communication network out of the call and text records and group meetings out of the Bluetooth interactions data.

Call and Text Messaging Data: We use the smartphone calls and texts exchanged by NetSense study participants from August 2011 to May 2013. Each communication record represents a call or text message, with the date, time, sender and receiver listed. We first divide the records time-wise into four snapshots, each snapshot representing one semester and then we create the corresponding four snapshots of the communication layer of the network.

Bluetooth Interactions Data: The Bluetooth interactions are recorded between pairs of study participants. We use only those interactions which are most likely to be face-to-face meetings by using a threshold of -65 dB on the RSSI value, which represents the signal strength of each interaction [13]. Similar to the call and text data, we divide the records into four snapshots. The records in each snapshot are used to find stable face-to-face groups in the corresponding semester.

Ego Network Surveys: Each student nominates up to 20 contacts at the beginning of every semester with whom she interacts the most frequently, via any interaction medium. However, typically only two to three of the 20 possible contacts listed are also study participants. We form a cognitive network out of the nominations made by students in the surveys. From our previous studies on this network [6] and [7], we know that the relationships underlying the network are very strong and distinguished from other kinds of relationships, with a lot of contacts and interactions between people connected by the nomination link. These links based on surveys form the cognitive layer of the network, which again is divided into four snapshots.

Node Attributes: Students participating in the NetSense study filled out a survey at the beginning of each semester. Survey questions were about the students family background, major pursued in Notre-Dame University, activities on campus, their views on various social issues, and their political inclinations. All attributes have multiple possible values out of which a student selects one. For example, students can select the value for their political views as either conservative or moderate or liberal. From many attributes available, for each student, we selected those, which were the most important for prediction of new links according to the reference [10]. These attributes are,

Student background

- Family income

Social and political views on:

- General political orientation (e.g. liberal/conservative)
- Opinion about legality of abortion
- Opinion about marijuana legalization
- Opinion about homosexuality and the legalization of gay marriage
- Views on racial equality and affirmative action

Habits and lifestyle:

- Drinking habits
- Time spent weekly on the following activities:
  - studying
  - partying
  - socializing
  - volunteering
  - campaigning for social causes
  - exercising.

III. Related Work

We use the NetSense Bluetooth network to identify face-to-face groups. This follows recent work done on group identification in social networks using non-obtrusive social sensors. References [14] and [12] introduce the method for identifying groups of people having a group meeting at a particular time, where the people interact with each other, as identified from Bluetooth interactions. The work explains why community detection is rendered unnecessary when groups can be identified from co-location. They make predictions on which individuals will join a group, identify core group members, and show why predicting the location of a group meeting is hard.

There has been some work on predicting group evolution and changing character of groups in social networks [1], [11]. The work in [1] performs predictions on large scale online groups based on social network websites, looking at how online group membership is influenced by the network structure surrounding the group and it's members. The work in [11] explores how opinions are influenced in a real-life network in an organization in which the groups of employees arose.

Going beyond this previous work, we introduce a method to identify stable-membership groups having face-to-face meetings over an elongated time period. Taking advantage of the attribute-rich nature of the NetSense data, we examine how personal attributes influence the formation of groups.

IV. Formation of Groups

We introduce a method to infer stable face-to-face groups using Bluetooth interactions between pairs of people. This method is significantly different from the approach used for identifying group meetings in [14] and [12] which identifies single group meetings happening at a particular time. In contrast, we aim to capture stable, long term groups of people, who meet regularly. We propose a hierarchical clustering-based method for identifying groups from records of time-stamped dyadic interactions.

The use of hierarchical clustering and small time snapshots for identifying groups has been well explained in [12]. In contrast, our method aggregates interactions over a long period of time and merges groups according to specific criteria. Our method for identifying stable groups aims to discover groups which have the following properties:

- Intersection: Two groups can be merged into a single group if there is at least a minimum level of intersection between the members of the two groups.
- Attendance: A group should include only members who have high attendance in the group meetings. While merging two groups, each of the members should have high attendance in all the combined group meetings of both groups.

A. Method to Identify Groups

Input: A set of interactions between pairs of people. Each interaction is associated with a time-stamp.

Output: Set of groups, each group consisting of a set of members.

An overview of our method and the intuition behind it:

1) We first identify potential group meetings from the Bluetooth interactions data. We observe that using small time-frames, like 10 minutes to one hour, to build snapshots of the network, we are able to observe several isolated small to medium sized connected components of nodes in each of the snapshots. When we use large

time-frames like 24 hours to one week to build snapshots of the network, we observe few but large sized connected components. Each connected component can be a potential group meeting. This has been observed in the work in [14]. We use a small time frame of 1 hour over the entire school semester to obtain several time intervals with such components existing in the network. We use these small-sized connected components to infer stable, long terms groups.

2) To infer stable groups from the identified connected components over several time intervals in the network snapshot for the given semester, we need to merge these connected components. Several components might contain people which do not happen to be regularly part of the group's meetings and may be accidentally included in a group meeting. On the other hand, certain group members who are otherwise regular members of the group might be absent from a few group meetings. These two cases necessitate a method which can merge the node connected components spread over time. We use a hierarchical clustering approach, which merges connected components based on their similarity measured by the number of common members and the number of group meetings of the components.

A more detailed description of this method is provided below: We start with a definitions of a few parameters and thresholds:

1) We define the *intersection*(g_1, g_2) between two groups g_1, g_2, as the number of members shared by both groups. The threshold for this intersection is referred to as $intThres$.
2) We define the *meetings-ratio*(m, g) for a node m and group g as the fraction of meetings a node m needs to attend to be considered part of the group g. We refer to the threshold for this value as $memThres$.
3) We define the *member-intersection*(g_1, g_2) between two groups, as a variant of the intersection score, in which the intersection of the two groups takes into account the members whose attendance for the particular group's meetings is above the $memThres$. We refer to the threshold corresponding to this score as $memIntThres$. We need a threshold on both the intersection and the member-intersection scores, since the intersection threshold ensures that only groups with a significant overlap of members are merged, and the member-intersection threshold ensures that only the members with a strong affiliation to the group are counted.

We now describe the algorithm in more detail:

1) **Identify connected components:** Choose a short time-frame, 10 minutes or one hour, to find multiple time intervals, each with at least one connected component in the entire network.
2) **Create groups:** For each time interval with connected components identify among them those components which have at least three people. Merge components which have the exact same members, across all the snapshots. Each of these components can be treated as a group and the set of all these components is denoted as $Groupset$.
3) **Merge groups:** Perform hierarchical clustering to merge groups in $Groupset$. Each group grp has a set of members defined by $member(grp)$.
   a) Iteratively, merge the pair of groups with the highest $memIntScore$, subject to the conditions listed below. A faster alternative to this approach is to iteratively merge a pair of groups which has the highest $memIntScore$ among all pairs in each iteration and stop when no pair of groups can be merged.
      - Merge the two groups $g1$ and $g2$ only if the $intersection(g1, g2) > intThres$.
      - Merge the two groups only if the $member-intersectionscore(g1, g2) > memIntThres$.
   b) Stop merging when no pair of groups satisfies the above conditions.

B. Properties of Identified Groups

Using the algorithm described in the previous subsection, we identify groups for each semester in the NetSense dataset. Table I lists the number of groups identified, the average number of people per group, the average number of meetings per group and the average attendance levels, i.e. the fraction of meetings attended by a member averaged over every group in each semester. Figure 1 shows the distribution of group membership size. We observe that an average group has about four members.

Most of the groups we obtain have three to six members, while a small number of groups are of a larger size. We experiment with different values of the thresholds, the $intThres$, $memIntThres$ and $memThres$ to obtain groups. Based on these experiments, we set the value of $intThres$ to 0.6, the value of $memIntThres$ to 0.5 and $memThres$ to 0.3. We observe that higher values of thresholds leads to smaller groups, however, with high attendance of members. This is because thresholds higher then chosen restrict the formation of larger groups and allow only triads, but those with very high amount of interactions among them. This is seen in Table II, we report properties of the groups found with threshold values increased for $intThres$ to 0.75, for $memIntThres$ to 0.75 and for $memThres$ to 0.5. More interesting are larger groups, which can be found by permitting some level of fluidity of members. On the other hand, thresholds lower than chosen lead to larger communities with lower member attendance in their meetings. Table III reports the properties of groups found with thresholds decreased, by setting $intThres$ to 0.3, $memIntThres$ to 0.3 and $memThres$ to 0.2. We observe that the groups discovered are larger in size and they hold many more meetings but with lower attendance. Obtaining a balance between these two scenarios, motivated our selection of threshold values we chose. While we have no ground truth to know if the groups which have been discovered are real,

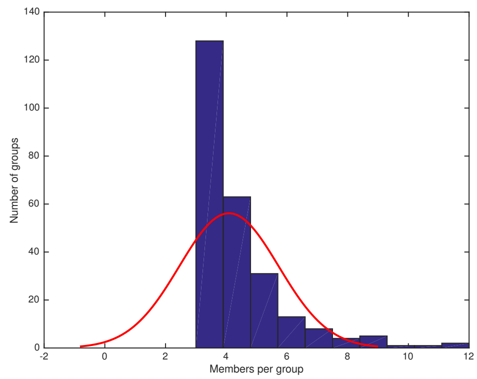

Fig. 1. Distribution of group sizes, most groups have three or four members. We also observe a significant number of groups with five to six members as well. The red line shows the best fit approximation to normal distribution for the data.

the values for the number of groups in a given population size, the number of groups per node, which is around five and the number of people per group, which is around four, makes sense intuitively.

TABLE I
PROPERTIES OF GROUPS FOUND WITH STANDARD THRESHOLDS ($intThres$ = 0.6, $memIntThres$ = 0.5, $memThres$ = 0.3).

Property	Semester			
	1	2	3	4
Number of people	189	185	171	153
Number of groups	256	269	149	100
Average group size	4.1	3.9	3.9	4.0
Number of meetings per group	177.9	224.5	205.1	132.8
Average member attendance	0.89	0.91	0.90	0.91

TABLE II
PROPERTIES OF GROUPS FOUND WITH HIGH THRESHOLDS ($intThres$ = 0.75, $memIntThres$ = 0.75, $memThres$ = 0.5).

Property	Semester			
	1	2	3	4
Number of people	189	185	171	153
Number of groups	223	251	134	93
Average group size	3.2	3.1	3.0	3.0
Number of meetings per group	102.1	103.6	81.3	70.1
Average member attendance	0.96	0.96	0.97	0.96

TABLE III
PROPERTIES OF GROUPS FOUND WITH LOW THRESHOLDS ($intThres$ = 0.3, $memIntThres$ = 0.3, $memThres$ = 0.2).

Property	Semester			
	1	2	3	4
Number of people	189	185	171	153
Number of groups	204	174	94	58
Average group size	5.9	5.5	5.6	5.8
Number of meetings per group	2240.1	4754.6	6520	6965.3
Average member attendance	0.73	0.75	0.74	0.74

V. INFLUENCE IN GROUPS AND COMMUNICATION DYADS

We explore what are the differences between group and communication contacts. We first find out what is the path length in the nomination network between pairs of people belonging to the same group and pairs of people connected by the communication edge. We then find out how much do pairs of people belonging to the same group and pairs of people connected by the communication edge differ on opinions and attribute values. Furthermore, we establish what is the difference in the selectivity of people towards their communication contacts and group contacts for attribute values and opinions.

A. Distance of Group Contacts in the Communication Network

From a previous work on the coevolution of the multiple layers in the NetSense network, we know that nomination and communication links are a significant indicator of relationship between people [6], [7]. Nomination links are sparser, have stronger communication and Bluetooth links and have higher survival rates. People connected by communication links are more numerous, are more likely to have nomination links between them, they also have stronger Bluetooth links and a higher survival rate compared to people which are connected only by Bluetooth links. So, in the hierarchy of strength of links, nomination links are the strongest, followed by communication links, followed further by Bluetooth interaction links.

We want to find out where do links joining people who belong to the same group lie in this hierarchy. To this end, we look at how far do group contacts lie from each other in the communication layer of the network. A short path would indicate a strong relationship. Intuitively, we expect there to be a path with length greater than one, between group contacts, since several groups would have mainly casual social or work related meetings. So, direct communication links may not exist between several pairs of members of groups. We first measure communication paths connecting pairs of people which belong to the same group and such paths between random people, and then we compare the results. As shown in Table IV, on average, pairs of people belonging to the same group tend to be two to three hops away from each other, while random contacts are about four hops away from each other in communication layer, the random paths are at least 40% longer than the in-group paths. The test of statistical significance of the result shows that the difference between path lengths connecting random and group contacts is statistically significant, with the p-value less than 0.01. The p-value is calculated from the Z-Score of the test statistic of the mean values of the path lengths [8].

Next, we measure the length of the paths connecting pairs of communication and group contacts in the nomination network. As shown in Table V, the communication contacts are closer to each other in the nomination network, than the group contacts are. We compute the statistical significance on the results as well, connecting communication and group contacts in the nomination layer is statistically significant, again with the p-value less than 0.01. Hence, these results demonstrate that while a group contact relationship is not as strong as a communication or a nomination relationship, it still has a significant amount of strength.

TABLE IV
DISTANCE BETWEEN GROUP CONTACTS IN THE COMMUNICATION LAYER.

Semester	Group Contacts	Random Node Pair
1	2.6	3.8
2	2.7	3.9
3	2.6	3.7
4	2.6	3.8

TABLE V
DISTANCE BETWEEN GROUP AND COMMUNICATION CONTACTS IN THE NOMINATION LAYER.

Semester	Communication Contacts	Group Contacts
1	1.2	3.6
2	1.2	3.5
3	1.3	3.7
4	1.8	3.8

B. The Difference in Agreement on Attribute Values for Group and Communication Contacts

We examine whether people are more likely to have common attribute values if they are linked by a communication edge and also belong to the same group. A higher level of agreement is indicative of a stronger homophily between people. The number of agreements is calculated by counting on how many of the 14 attributes listed in the Data section a pair of people agrees. We compare the average number of attribute agreements between people belonging to the same group and people connected by communication links. We also compare these numbers to the average number of agreements between pairs of people which are not connected by any of these relationships. We find that pairs of the last type tend to have the least number of agreements, followed by the pairs of people with only a group relationship between them and finally followed by the pairs of people with a communication relationship. Table VI lists these differences. This observation again reaffirms that although group relationships are not as strong as communication relationships, in terms of attribute value agreement, there is still a significant amount of selection which happens when groups are formed or when nodes join groups. This suggests that collectively the members of a group exert some influence over the other members, so unfitting members either quit or adapt more attribute values and opinions possibly more agreeable to the majority of other members.

C. Bias on Attribute Values for Communication and Group Contacts

Here, we examine if there is any selectivity in forming or joining a group based on the attribute values which people in the group already possess. We are particularly interested in finding if people have a strong bias for groups with attribute

TABLE VI
NUMBER OF ATTRIBUTE AGREEMENTS AMONG GROUP MEMBERS, BETWEEN PEOPLE CONNECTED BY THE COMMUNICATION LINKS AND RANDOM PAIRS OF PEOPLE.

Semester	Communication Contacts	Group Contacts	Random Contacts
1	8.1	7.2	5.1
2	8.0	7.1	5.3
3	8.3	7.2	5.5
4	8.5	7.4	5.4

values similar to their own. For this purpose, we compare the bias that people have for their communication contacts for the same attribute value as that of the node for which we measure the bias, with the bias that people have for the groups they join. To measure the latter bias, we first count the relevant group members that have the same attribute value as that of the node for which biased is measured. Then we calculate the Z-scores for this node for having that actual number of contacts with the same attribute value as the node. This score is based on the number of people expected to hold this attribute value, based on the number of people with the particular attribute value in the entire network. This result represents the level of selectivity the node has for the particular attribute value. These Z-Scores are calculated for both people connected by communication links and for members of the same group as,

$$Z - Score(val, n) = \frac{x - \mu}{\sigma} \quad (1)$$

where, x is the actual number of contacts of node n with the attribute value val, μ is the expected number of contacts with the value v in the network, σ is the standard deviation. The values for the Z-Score range between -3.4 and $+3.4$. A value close to zero indicates lower bias, and a negative value indicates a negative bias. We normalize the Z-Score to linear score in the range -1 and $+1$. Table VII lists the average selectivity values for communication links and among members of the same group. We observe that for several attributes members of the same group have a statistically significant lower bias, at the significance level of 0.05 then people connected by communication links. These attributes are listed in Table VII in bold italics. We know from the previous sub-section, that people have higher level of attribute value agreement across their communication links than among the members of the same group. These attributes exclude most of the lifestyle related attributes (such as time spend on different activities). Yet, the bias of each node for most of the attributes is higher for members of the same groups than among pairs of random people. With this observation, we know that people have some selectivity for which groups they would join, based on the views and opinions of members in the group. Although this bias is not as strong as the bias they have for forming communication links, it is still quite significant and thus influences which groups people would join.

TABLE VII
BIAS AS MEASURED BY THE NORMALIZED Z-SCORE OF PEOPLE FOR THE NODE'S ATTRIBUTE VALUE. ATTRIBUTES WITH STATISTICALLY SIGNIFICANT DIFFERENCES ARE SHOWN IN BOLD ITALICS.

Attribute	Communication Link	Group Link	P-value
Political views	0.16	0.12	0.058
Parental income	***0.18***	***0.07***	***0.001***
Drinking habits	***0.12***	***0.06***	***0.008***
Views on abortion	***0.17***	***0.12***	***0.042***
Views on gay marriage legalization	0.10	0.06	0.123
Views on homosexuality	***0.12***	***0.04***	***0.002***
Views on marijuana legalization	0.09	0.04	0.063
Time spent on volunteering	0.04	0.03	0.624
Time spent on exercising	0.06	0.05	0.600
Time spent on studying	***0.15***	***0.07***	***0.004***
Time spent on partying	0.03	0.03	1.000
Time spent on university clubs	0.10	0.08	0.425
Time spent on socializing	***0.10***	***0.03***	***0.021***
Time spent on camping	0.01	0.05	0.143

VI. POLARIZATION AND DIFFERENCES AMONG GROUPS

Now that we have found that people are selective while forming groups, we would like to find out if there exists some observable kind of needs to be primary fulfilled by group membership. We found that we can distinguish between groups in which the focus is mainly on social needs of their members and groups whose primary focus is on school performance. We will refer to the former type of groups as social groups and to the latter type as work groups. We aim then at answering the question if the type of a group can tell us how strong is the influence of a group on its members.

The group type can be inferred from the behavioral properties of the group, and from influence of the attribute values in the group. While we cannot infer the exact purpose of every group, we were successful in distinguishing between work groups and social groups. In addition to this, we establish to what extent groups can be clustered in terms of homogeneity of group members with respect to their attribute values.

First, we find what type each group is: whether the group meetings are more likely to enable members to work together or to socialize together. To this end, we try to cluster groups based on the temporal features of the group's meetings. We then check if the groups with these two different purposes have different levels of selectivity. This enables us to verify if groups are selective over not only individual attribute values, but also over multiple related attribute values.

A. Inferring Group Type

To identify a type of every group in the network, we cluster each group based on the timings of its meetings. We use the percentage of group meetings which happen on weekdays, during the day versus those which happen on weekends and weekday evenings as a feature for k-means clustering. With k=2, we obtain two well-defined clusters. One of the clusters contains groups with meetings happening predominantly on weekdays during the daytime. The other cluster contains groups with meetings evenly distributed across weekday daytimes and weekends and weekday evenings. The groups in the first cluster can be inferred to have meetings at which members work and study together. On the other hand, groups in the second cluster are likely to be of a more social nature, where group meetings happen in both work-study hours as well as on weekends. On an average, 30% of groups are classified as social while the remaining 70% are work groups. This finding is consistent with some studies on groups [14].

B. Relation Between Bias and Type of a Group

We want to know if the different kinds of groups we previously identified, the work groups and the social groups, have different levels of selectivity when it comes to attribute values and opinions. It seems reasonable to expect that work groups are likely to have lower selectivity with this regard than social groups do. To verify this intuition, for each group of work type and of social type, we measure the average selectivity for attribute values chosen for analysis in the subsection V-B.

As before, the selectivity is measured by calculating the Z-Score for the and normalizing it to the range -1 to +1. As shown in Table VIII work groups are significantly statistically less selective than social groups for several attributes. Those attributes are both, opinions and lifestyle choices, including parental income, drinking habits, views on homosexuality, marijuana legalization and time spent studying. The difference is not significant for the rest of the attributes.

TABLE VIII
SELECTIVITY OF WORK AND SOCIAL GROUPS. ATTRIBUTES WITH STATISTICALLY SIGNIFICANT DIFFERENCES ARE SHOWN IN BOLD ITALICS.

Attribute	Work Groups	Social Groups	P-value
Political views	0.74	0.79	0.274
Parental income	***0.20***	***0.26***	***0.032***
Drinking habits	***0.36***	***0.55***	***0.002***
Views on abortion	0.70	0.75	0.183
Views on gay marriage legalization	0.64	0.67	0.365
Views on homosexuality	***0.06***	***0.42***	***0.001***
Views on marijuana legalization	***0.52***	***0.76***	***0.001***
Time spent on volunteering	0.32	0.35	0.123
Time spent on exercising	0.62	0.65	0.095
Time spent on studying	***0.62***	***0.69***	***0.036***
Time spent on partying	0.43	0.46	0.142
Time spent on university clubs	0.52	0.57	0.074
Time spent on socializing	0.56	0.61	0.148
Time spent on camping	0.72	0.75	0.356

C. Attribute based Clustering of Groups

We have found that groups are selective on several attributes individually. Now we examine if groups are selective for a set of related attribute values. Influence by groups might possibly happen through several attributes at a time. For example, attribute values like liberal political orientation, a positive view of homosexuality and a favorable view of abortion might all be positively correlated among people. To examine the selectivity of groups on several attributes at a time, we cluster them by the percentage of group members with various attribute values of each group. The attributes we use for clustering groups are: parental income, political views, views of homosexuality, marijuana, gay marriages and abortion legalization. The features for clustering are the percentage of group members for each possible value for each of the attributes. We cluster groups using k-means clustering. We experiment with k=2 and with k=3 clustering. Table IX and Table X list the most prevalent attribute values in each cluster. With k=2, we are able to get two clusters of groups, each distinct from each other in terms of the dominant attribute value in each group. The first cluster consists of groups with most members with strongly conservative political leaning who are against gay marriages and homosexuality; these members typically have high-income parents. The second cluster has groups whose members are leaning towards the liberal views. With k=3, as seen in Table X, we observe that the cluster with the liberal views splits into two clusters, one of the clusters has groups whose members have largely moderately liberal views, and the other cluster has groups with members having strongly liberal views. There is no change to the cluster with groups having members with conservative views. The clustering of groups based on attribute values shows that the formation of groups can be potentially influenced by several attributes. Each attribute may not act individually, yet several attributes may act collectively to influence who the members of a group might be.

TABLE IX
CLUSTERING OF GROUPS WITH K=2

Attribute	Cluster 1	Cluster 2
Views on abortion	Against	Largely for
Views on gay marriage	Moderately for	Largely for
Views on homosexuality	Against	For
Views on marijuana legalization	For	Equally distributed
Family Income	Mostly rich, few poor	Rich and middle class
Political Views	Mostly conservative	Evenly distributed, liberal bias

The observation that groups within the network can be neatly split into clusters indicates a strong underlying influence process among people, which leads to homogeneous groups, and possibly opinion influence among members.

VII. CONCLUSIONS AND DISCUSSION

We propose a method to identify stable, interacting groups of people from Bluetooth interaction records. First we identify ways in which group relationships are different from communication relationships: while the former show a significant amount of selectivity, which influences how groups are formed, they are not as selective as the latter. We then find how the type of a group can be inferred from the temporal features of group meetings and how this type relates to the selectivity of groups. Finally, we discuss how multiple related attribute values influence the formation of groups, indicating that selectivity on some opinions and attribute values during group formation imposes polarization among groups while others are not statistically significantly selective. We also observe a strong difference of selectivity for group membership between work groups and social groups for four attributes. Those attributes are parental income, drinking habits, views on homosexuality and views on marijuana legalization. This is an interesting but not surprising observation since all four attributes impact strongly what kind of leisure activities the members might enjoy. Indirectly, this observation validates our method of identifying groups and their types.

TABLE X
CLUSTERING OF GROUPS WITH K=3

Attribute	Cluster 1	Cluster 2	Cluster 3
Views on abortion	Against	Largely for	Strongly for
Views on gay marriage	Moderately for	Largely for	Strongly for
Views on homosexuality	Against	Largely for	Strongly for
Views on Marijuana legalization	For	Equally distributed	Weak bias for against
Family income	Mostly rich, few poor	Rich, few middle class	Largely middle class
Political views	Conservative	Largely liberal	Strongly liberal

ACKNOWLEDGMENT

This work was supported in part by the Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA), and by the Office of Naval Research (ONR) Grant No. N00014-15-1-2640. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the Army Research Laboratory or the U.S. Government.

REFERENCES

[1] L., Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, X., "Group formation in large social networks: membership, growth, and evolution," *Proc. 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006 pp. 4454.

[2] L .C. Freeman, "The sociological concept of" group": An empirical test of two models." *American Journal of Sociology*, **98** 1992, pp. 152–166.

[3] G.A. Fine and B. Harrington, "Tiny publics: Small groups and civil society." *Sociological Theory*, 2004, pp. 341-356.

[4] M.T. Rivera, S.B. Soderstrom and B. Uzzi, "Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms." *Annual Review of Sociology*, 2010, pp.91-115.

[5] A. Bahulkar, B. K. Szymanski, O. Lizardo, Y. Dong, Y., and N. Chawla, "Analysis of link formation, persistence and dissolution in NetSense data," *Proc. 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 1197–1204.

[6] A. Bahulkar, B. K. Szymanski, K. Chan, and O. Lizardo, "Co-evolution of two networks representing different social relations in NetSense," *Proc. International Workshop on Complex Networks and their Applications*, Studies in Computational Intelligence Series, **693**, 2016, Springer, pp. 423–434.
[7] A. Bahulkar, B. K. Szymanski, K. Chan and O. Lizardo, "Coevolution of a multilayer node-aligned network of which layers represent different social relations," *Computational Social Networks*, 2017.
[8] D. C. Montgomery, and G. C. Runger, "Applied statistics and probability for engineers," Hoboken, NJ: Wiley, 2007, pp. 294 – 296.
[9] J.M. McPherson and L. Smith-Lovin, "Homophily in voluntary organizations: Status distance and the composition of face-to-face groups," *American Sociological Review*, **52(3)**, 1987, pp. 370-379.
[10] A. Bahulkar, B. K. Szymanski, K. Chan and O. Lizardo, "Influence of Personal Preferences on Link Dynamics in Social Networks," *Complexity*, 2017.
[11] A. Lampinen, S. Tamminen, and A. Oulasvirta, "All my people right here, right now: management of group co-presence on a social networking site," *Proc. ACM 2009 International Conference on Supporting Group Work*, 2009, pp. 281-290.
[12] V. Sekara, A. Stopczynski, S. Lehmann, "The fundamental structures of dynamic social networks," Proceedings of the National Academy of Sciences, **113(36)**, 2016, pp. 9977 - 9982.
[13] L. Shu, J. Yingxin, and A. Striegel, "Face-to-Face Proximity Estimation Using Bluetooth On Smartphones," *IEEE Transactions on Mobile Computing*, **13(4)**, 2003, pp. 811 - 823.
[14] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. Larsen, and S. Lehmann, "Measuring large-scale social networks with high resolution," *PloS ONE*, **9**: e95978, 2014.
[15] Hu, X., Meng, L., and Striegel, A. D., "Evaluating the raw potential for device-to-device caching via co-location," *Procedia Computer Science*, **34**, 2014, pp. 376-383.
[16] A. Striegel, S. Liu, L. Meng, C. Poellabauer, D. Hachen, and O. Lizardo, "Lessons learned from the netsense smartphone study," *ACM SIGCOMM Computer Communication Review*, **43(4)**, 2013, pp. 51–56.
[17] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *Journal of Mathematical Sociology*, **34**, 1990, pp. 193–206.