

Determining Infectious Disease Positivity Rate Over Interaction Rate Through Analysis of Collocation Data

Yiyang Wang

Faculty Member #1:

Printed: Gregory Abowd

Signature: 
3B96123051554D8...

Faculty Member #2:

Printed: Thomas Ploetz

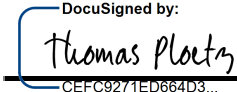
Signature: 
CEFC9271ED864D3...

Table of Contents

Abstract.....	2
Introduction.....	3
Literature Review.....	6
Methodology.....	10
Data Collection and Manipulation.....	10
Measures.....	14
Results and Discussion.....	15
Results.....	15
Discussion.....	17
Conclusion.....	18
References.....	20

Abstract

Knowing and understanding the flow of infectious disease within a community is very important, as it can effectively aid community administrations in planning their actions to control the situations. In this study, I investigated the correlation between collocation rate and COVID positivity rate within each building pair using Wifi data and school daily COVID case reports. I used collocation bipartite graphs to generate occupancy features, such as occupancy count and occupancy duration, and then input these features into our correlation model. I then used the COVID positivity rate as the output of the correlation model. Using the model, I am able to find a weak linear correlation between the collocation rate and the COVID positivity rate. The correlation is stronger in places with more students in and out on a daily basis, such as libraries and student centers. Using this insight, the school administration could advise students who visited these places to get tested when there was a COVID outbreak in these areas. A similar approach could also be adopted to investigate the correlation between collocation rate and infectious positivity rate on other types of infectious diseases in the future.

Introduction

A crowd is a collection of people who happen to be in the same place at the same time. A collective or crowd behavior is the relatively spontaneous and relatively unstructured behavior by large numbers of individuals acting with or being influenced by other individuals.[1] Detecting social groups and understanding crowd behaviors in a large social context is very important and valuable, as it could provide insights for event organizing, marketing, urban planning, healthcare, and many other areas.[2,3] One example of the usage of group information that was debated over its morality is marketing products targeting specific groups at different locations. Another example is to use group information for resource allocation. By understanding the social group behavior across a university campus, the school administration could have a better planning of school buses.

A traditional way to detect social groups is to place cameras in public areas, but it suffers from issues like coverage and costs. Nowadays the use of mobile devices is increasing exponentially and detection using data collected from mobile devices, such as app usage, messaging and WiFi connection points is easier and more accurate. Although collecting data from mobile devices to monitor crowds has become a hot research topic, not enough attention is brought to the thorough and effective analysis of the collected data.[3]

Previous relevant studies proposed several methods for analyzing data collected through passive WiFi monitoring on mobile devices. These methods do not require additional apps or hardware besides your mobile device to collect data. It will collect the information from where you connect to WiFi using your mobile device. Y. Zhou and etc.

suggested a data analysis framework with the usage of statistics, visualization, and unsupervised machine learning to analyze patterns of crowds by days and locations in a large social event.[3] Prior work in my lab by Harish Haresamudram, etc. suggested a novel self-supervised pre-training model that analyzes human activity through data collected from wearable devices, such as smartwatches. The inputs of this pre-training model are unlabeled movement data and the outputs are movement data with reconstruction at timestamps in a defined format. Then we could use this new movement data to recognize and classify activities, such as running and standing.[4]

Previous studies proposed methods to analyze social groups for one location at a time, however, in my work, a model is proposed to analyze social groups interaction detected at different locations across multiple time periods to train and predict the COVID positivity rate based on the interaction rate. Instead of focusing on one building location each time, I focus on the collocation time and COVID positivity rate in pairs of buildings to infer how the infected COVID cases flow across social groups on campus and to understand how different social groups interact across campus.

COVID positivity rate is defined as the percentage of residents who got COVID and live at a specific residence house on campus over the total number of residents who got COVID on campus. By applying clustering algorithms at each location with passive WiFi sensing data collected from the Georgia Tech campus, the ambition is to map the clusters and investigate the interaction amount between the social groups at each location. Then using the COVID positivity rate at each location, we try to predict the percentage of positivity cases over the total number of interactions.

Through this study, the goal is to not only get a more accurate social group detection result, but also an understanding of the correlation between social group interaction, specifically their collocation duration, and the spread of infectious disease. Understanding social group interactions and disease spreading could help with campus resource allocation and better campus-related planning in the future. For example, a notification could be sent to students to suggest they get tested based on our prediction. To achieve this goal, the first step would be to build features based on collocation data. Then COVID positivity rate on each day in each building on campus is calculated and I used this data to calculate the total positivity rate for selected building pairs. The features serve as inputs and the COVID positivity rates serve as outputs of the correlation model to be constructed in this study. Lastly, using this correlation model, predictions are made using features on randomly selected days and then the results' significance will be calculated based on actual COVID positivity rates.

Literature Review

Research on analyzing crowd behavior without additional software and hardware is gaining popularity and having crowd behavior information could help us in many areas. Wireless communication is constantly developing, and mobile technologies are prevalent in our daily life more and more, so mobile social networking is rising and becoming a hot research topic [5,6]. As the market for mobile devices grows, multiple different systems are developed to complement the use of smart mobile devices. With the development of these different platforms, such as Android and IOS, it is hard for companies to produce services that work across all of the platforms and therefore they limit their service to certain platforms [5].

Yet information gained from these platforms is very valuable. It is shown that the information gained from mobile social interaction, specifically the detection of co-location, which is a group of people located at the same place at the same time, could indicate some extent of social interactions in real-life situations [7]. Using this information, we could analyze the shape of social interactions, such as the frequency, places and time people interact. Thus, investigating the shape of social interactions and identifying the social groups from mobile technologies could help us improve existing platforms and shape various community networks enhancing different needs.

Co-location data provides a proxy for real-life interactions. A popular way to detect co-location is through collecting WiFi data because WiFi data can be collected on all devices and platforms. Vanderhulst et al. proposed the use of WiFi probes to periodically capture the signals from mobile devices and WiFi access points to detect human interactions. They utilize sociology concepts and define a high-accuracy model

that filters those in the same location by having short-lived, spontaneous social interactions with each other [8]. However, a large number of people could be using the same WiFi access point and one WiFi access point may be able to cover a wide area. Thus, it is challenging to filter the different small groups within the area in the span of a WiFi access point.

Multiple methods are used to further break down the groups at the same WiFi access point. Hong et al. proposed a more detailed part of the WiFi data that could be used as an indicator for social interaction, the “Receive Signal Strength Indicators (RSSI)”. RSSI could indicate people’s distance from a WiFi access point, therefore providing higher accuracy than just identifying the locations of different WiFi access points. The researchers created the SocialProbe system to passively monitor WiFi probe requests sent from smartphones and get RSSI from the data [9]. Jon Baker and Christos Efstratiou took a different approach and proposed a system called Next2Me to analyze both the WiFi signals and audio signals from microphones on mobile devices to distinguish social group interactions with close proximity, as close as “a few meters from each other”. The utilization of sound fingerprinting in the study could achieve a high precision even in noisy environments and the addition of sound detection is not using battery significantly [7].

While some research in this field pertains to collecting WiFi data, other research seeks to analyze it. In A. Mtibaa et al.’s study, a novel Social Aware Cluster Based Localization algorithm (SAC-Loc) is proposed. The algorithm helps to improve clustering and avoids unintentional splitting of a group due to wireless scanning limitations or joining groups with temporary proximity [11]. In addition, Y. Zhou et al. proposed an

effective data analysis framework – collect data, analyze spatial patterns, apply clustering algorithms and combine with time to produce spatiotemporal patterns [3]. These analysis algorithms could help accompany the methods to break down WiFi data for producing a more precise model for categorizing social groups. In my study, I will use co-location data processed by these analysis algorithms.

The above studies provide an excellent foundation for producing a model that utilizes WiFi data to analyze social groups, but the method and framework used in these studies could be improved with further larger-scale studies. Currently, although there are studies on improving the analysis of the WiFi data, “not enough attention has been paid to the thorough analysis and mining of collected data” [3] and there are very limited studies trying to improve the efficiency of the WiFi data analysis algorithms. This study brings attention to a further analysis of Wifi data collected. Furthermore, there are relevant sociology studies specifically targeting college students’ social groups [12], but not many previous studies focus on analyzing college students’ social group detection using WiFi colocation. This study is going to analyze college student social groups using Wifi colocation data.

A data analysis model based on the above studies is being proposed, but will be utilizing large-scale college students’ mobile WiFi data to generate a precise and applicable analysis of college students’ social group. Then through analyzing the interaction amount within the social groups, we take data on infectious disease positivity rate and determine the positivity rate based on interaction rate. The results of such a study would be useful to the college administration and aid them in providing a better

and healthier learning community for college students. Details on how the data is collected and how the model is built are discussed in the following methodology section.

Methodology

We looked at methods used in previous studies and adopted and modified the following method procedures.

Data Collection and Manipulation

In this study, Wifi access points data are collected from Georgia Tech's on-campus Wifi data access points. Researchers are Georgia Tech students and faculty members and they need to log onto Georgia Tech Virtual Private Network(VPN) to access these datasets. Researchers all agreed to not download the data to the local computer or leak personal information in the datasets. WiFi data is passively collected and participants are represented with anonymous user ids. Every unnamed participant in the study data is a resident of an on-campus location and each of them has a unique user id. This is used to determine user pairs for colocation.

Colocation data is calculated from the Wifi access datasets. All of the files could be opened using Jupyter Notebook. Jupyter Notebook is an open-source web application that allows users to edit and run code and add graph visualizations on their browsers. Python and Python modules, such as Pandas and Matplotlib, are used in the study.

Every combination of any two buildings on campus is considered a building pair. In each building pair, every two users' colocation data is determined. As shown in Table 1 below, a table is generated to store the colocation data each day for every building pair, where each row contains the colocation data for each user group. The colocation

data includes users' unique ids, unique building ids, and length of time together. The tables for colocation data on each day are stored as a csv file.

	end_time	sem_loc	start_time	users	duration	b_id
0	2020-11-11 17:51:00	077:g252	2020-11-11 17:46:00	[1266326, 1039013]	5.0	077
1	2020-11-11 21:57:00	077:g252	2020-11-11 21:56:00	[1563570, 1677732]	1.0	077
2	2020-11-11 21:59:00	077:g252	2020-11-11 21:57:00	[1563570, 1677732]	2.0	077
0	2020-11-11 09:29:00	077:g290	2020-11-11 09:24:00	[1563352, 1039393]	5.0	077
1	2020-11-11 09:39:00	077:g290	2020-11-11 09:31:00	[1563352, 1021687]	8.0	077
2	2020-11-11 09:44:00	077:g290	2020-11-11 09:39:00	[1563352, 1021687, 1052464]	5.0	077
3	2020-11-11 09:54:00	077:g290	2020-11-11 09:44:00	[1563352, 1021687, 1052464, 1021801]	10.0	077
4	2020-11-11 10:08:00	077:g290	2020-11-11 09:54:00	[1563352, 1021687, 1052464, 1021801]	14.0	077
5	2020-11-11 10:15:00	077:g290	2020-11-11 10:08:00	[1021687, 1052464, 1021801]	7.0	077
6	2020-11-11 10:21:00	077:g290	2020-11-11 10:15:00	[1021687, 1052464, 1021801, 1683305]	6.0	077
7	2020-11-11 10:27:00	077:g290	2020-11-11 10:21:00	[1021687, 1052464, 1683305]	6.0	077
8	2020-11-11 11:09:00	077:g290	2020-11-11 11:07:00	[1022525, 1037422]	2.0	077
9	2020-11-11 11:21:00	077:g290	2020-11-11 11:09:00	[1022525, 1037422, 1030996]	12.0	077
10	2020-11-11 11:24:00	077:g290	2020-11-11 11:21:00	[1022525, 1037422, 1030996, 1022626]	3.0	077
11	2020-11-11 11:28:00	077:g290	2020-11-11 11:24:00	[1022525, 1037422, 1030996, 1022626, 1039393]	4.0	077
12	2020-11-11 11:29:00	077:g290	2020-11-11 11:28:00	[1022525, 1037422, 1030996, 1022626]	1.0	077
13	2020-11-11 11:32:00	077:g290	2020-11-11 11:29:00	[1022525, 1037422, 1030996, 1022626, 1039148]	3.0	077
14	2020-11-11 11:34:00	077:g290	2020-11-11 11:32:00	[1022525, 1037422, 1030996, 1039148]	2.0	077

Table 1

Then colocation data is then mapped onto a bipartite graph. A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V . In this study, the bipartite graph contains users and wifi access points as its two disjoint and independent sets. Bipartite graphs are used because every edge is unique, therefore in this case, every user group's collocation data is guaranteed to be unique.

Using this bipartite graph, three occupancy features are calculated: occupancy count, occupancy period and occupancy duration. Occupancy count is the unique

number of visitors by some dwelling time > threshold every day. Occupancy period is the timestamp when people arrive and leave. Occupancy duration is the average dwelling time.

A function is created to calculate each occupancy feature by traversing through the vertices and edges in the bipartite graph, as shown in Figure 1, Figure 2, and Figure 3 below. Firstly, the bipartite graph is read in from the stored service location of the data into each function using the python NetworkX library. Then, the function traverses through every edge in the bipartite graph. User data and location data are assigned using the two nodes of each edge. Different computations are performed afterward using the user and location data to calculate occupancy features.

```
def getOccupancyCount(date_d, threshold):
    out_row = {}
    out_row['date'] = date_d

    BG_path = SOURCE_DIR + 'graph_dwelling_segments_' + date_d.strftime('%Y%m%d') + '.p'
    BG = nx.read_gpickle(BG_path)
    usr_nodes = {n for n, d in BG.nodes(data=True) if d['bipartite']==0}

    occupancy_count = 0

    daily_edges = [(u, v, d) for u,v,d in BG.edges(data=True)]

    for edge in daily_edges:
        usr = edge[0]
        if usr not in usr_nodes:
            usr = edge[1]
            if edge[2]['dur'] > threshold:
                occupancy_count = occupancy_count + 1

    out_row['occupancy_count'] = occupancy_count

    print("Done", BG_path)

    return out_row
```

Figure 1

```

def getOccupancyPeriod(date_d):
    out_row = {}
    out_row['date'] = date_d

    BG_path = SOURCE_DIR + 'graph_dwelling_segments_' + date_d.strftime('%Y%m%d') + '.p'
    BG = nx.read_gpickle(BG_path)

    occupancy_period = []

    daily_edges = [(u, v, d) for u,v,d in BG.edges(data=True)]

    for edge in daily_edges:
        occupancy_period.append((edge[2]['start_t'], edge[2]['end_t']))

    out_row['occupancy_period'] = occupancy_period

    print("Done", BG_path)

    return out_row

```

Figure 2

```

def getOccupancyDuration(date_d):
    out_row = {}
    out_row['date'] = date_d

    BG_path = SOURCE_DIR + 'graph_dwelling_segments_' + date_d.strftime('%Y%m%d') + '.p'
    BG = nx.read_gpickle(BG_path)

    edge_count = 0
    occupancy_duration = 0

    daily_edges = [(u, v, d) for u,v,d in BG.edges(data=True)]

    for edge in daily_edges:
        occupancy_duration = occupancy_duration + edge[2]['dur']
        edge_count = edge_count + 1

    out_row['occupancy_duration'] = occupancy_duration / edge_count

    print("Done", BG_path)

    return out_row

```

Figure 3

Measures

Colocation data is read from the corresponding csv files using Pandas module in Python on Jupyter notebook. The total length of time students spend together at each building pair is calculated using occupancy features. For each building pair, this total length of time is added. This could be achieved through filtering the unique user ids for each student pair and the unique building ids for each building pair. Then a colocation rate is calculated by $collocation\ rate = \frac{total\ collocation\ time\ at\ one\ building\ pair}{total\ collocation\ period\ at\ all\ building\ pairs}$. Using the dataset of COVID positive number, a positivity rate for the residents in a building is calculated by $positivity\ rate = \frac{positive\ case\ number\ at\ a\ building}{total\ positive\ case\ number\ on\ campus}$. Five places with highest positivity rate and five places with lowest positivity rate are chosen. The colocation rate and positivity rate are both normalized to a number between 0 and 1.

It is anticipated that this data analysis model will show that higher colocation rate is positively correlated with higher positive rate. The details of our study results are discussed in the next section, results and discussion.

Results and Discussion

Results

Using the data collection and manipulation and measures in the methods section, I am able to obtain the following results.

In the results, there is a mild correlation between the collocation rate of a building and the COVID positivity rate of the residents living in that building. The significance of the results still needs to be calculated.

First, a table is generated with occupancy features. Occupancy features are obtained by calling each function with an input of a certain day¹. Then collocation rate is calculated using the occupancy count, occupancy period, and occupancy duration. A two-dimensional scatter plot is used to show the correlation between the collocation rate and the COVID positivity rate at these ten locations. In Figure 7 below, it maps the collocation duration versus COVID positivity rate in the five locations with the highest COVID positivity rate on campus. It is shown that there is some positive correlation between collocation duration and COVID positive rate, but there might be places with anomalies. For example, local residents may be in contact with people who are not Georgia Tech residents such as their family members and their colocation duration would not be reflected.

¹

https://github.gatech.edu/ubicomp/contact_tracing/blob/diana-predict-positivity/code/notebooks/occupancy_features.py

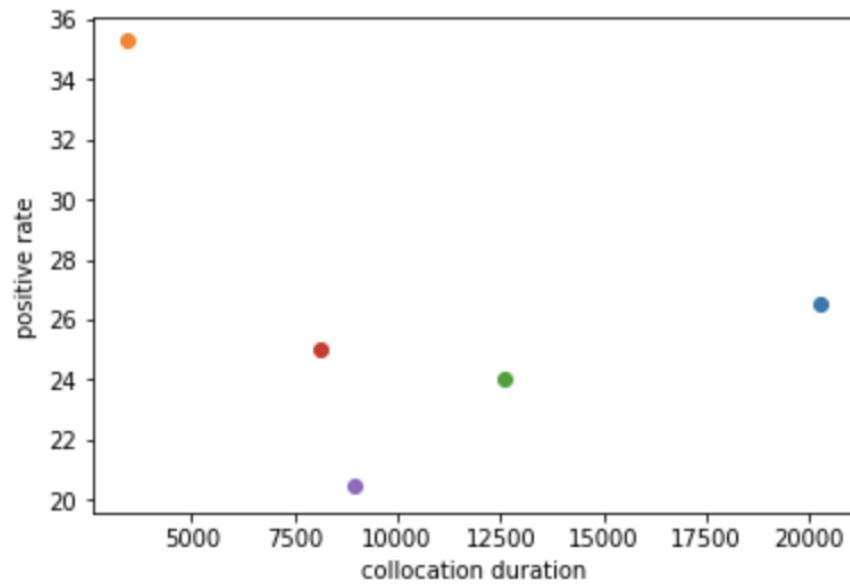


Figure 7

The significance of the results is discussed further in the next discussion section.

Discussion

Similar to the approach used in Mehrab et al.'s paper "High-resolution proximity statistics as an early warning for US universities reopening during COVID-19" [13], this approach utilized occupancy features and COVID infection data to analyze the correlation between collocation duration and COVID positivity rate. However, in Mehrab et al.'s study, they focused on university social groups' collocation with the surrounding communities. In this study, it is focused on university social groups' interactions on campus through analyzing data with building pairs. Our results share a similar trend as in that article, but there is only COVID positivity data and collocation data on one campus in this study. Thus, this study is focused more on the flow of COVID positive cases within the campus rather than in and out of campus. In the next step, the study could be focused on analyzing the collocation data on a day several days prior to a day with some positive cases, so that the model could use collocation duration to predict a spike in cases several days later.

From these aspects, this study could be concluded as the following.

Conclusion

This data analysis model maps the correlation between colocation rate and COVID positivity rate using data collected from Georgia Tech's main campus. Using a similar approach, this model could be easily modified to fit the data collected from other school campuses. By having this data analysis model to denote and analyze the correlation between colocation rate and COVID positivity rate, the school administration could better plan our school bus system and other actions during the pandemic to build a safer environment for students and faculties. For example, it is predicted that the student center and library would have a higher correlation between colocation rate and COVID positivity rate. School administration could hold higher sanitary standards or stricter rules at such places and places where the positivity rate is usually higher than elsewhere.

COVID is not the only infectious disease globally and historically. This model is applicable to other infectious diseases as well. The disease positivity rate could be calculated for other infectious diseases. A similar model could be generated to find the correlation between colocation data and other infectious diseases.

However, this model currently only selects two buildings as a building pair. The shown flow of the positivity rate over colocation rate across the whole campus is limited. If in future studies, a larger chain of buildings could be included for each model, the model could show a clearer representation of the correlation and possibly a flow of how the disease was spread. Having this analysis, school administration could predict the future transmission of the disease on campus and could have better prevention of the spread. For example, given a common route for disease spread, if one positive case

occurs at one location, the school administration could ask students living in other buildings at this route to test themselves. Currently, it is hard to perform the analysis on all user combinations if more buildings are included in each building pair because the Georgia Tech campus has a large number of residents and it is difficult to find the overlapping resident pairs from all of these combinations for more buildings. Thus, it would be a great addition to the model if more buildings are included in each building pair.

References

- [1] Barkan, S., 2012. Sociology: Comprehensive Edition. Chapter “Deviance, Crime, and Social Control”. 10th ed.
- [2] J. Shen, J. Cao and X. Liu, "BaG: Behavior-aware Group Detection in Crowded Urban Spaces using WiFi Probes," in IEEE Transactions on Mobile Computing, doi: 10.1109/TMC.2020.2999491.
- [3] Y. Zhou, B. P. L. Lau, Z. Koh, C. Yuen and B. K. K. Ng, "Understanding Crowd Behaviors in a Social Event by Passive WiFi Sensing and Data Mining," in IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4442-4454, May 2020, doi: 10.1109/JIOT.2020.2972062.
- [4] Harish Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick L. Grady, Irfan Essa, Judy Hoffman, and Thomas Plötz. 2020. Masked reconstruction based self-supervision for human activity recognition. In *Proceedings of the 2020 International Symposium on Wearable Computers* (*ISWC '20*). Association for Computing Machinery, New York, NY, USA, 45–49. DOI:<https://doi.org/10.1145/3410531.3414306>
- [5] Garcia-Barrios, Victor & Qerkini, Korab & Safran, Christian. (2009). What students really need beyond learning content: Ubiquitous shared-connectivity services to foster learning communities on the campus. 10.1145/1621841.1621869.
- [6] Z. Yu, Y. Liang, B. Xu, Y. Yang and B. Guo, "Towards a Smart Campus with Mobile Social Networking," 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing, Dalian, 2011, pp. 162-169, doi: 10.1109/iThings/CPSCoM.2011.55.
- [7] Jon Baker and Christos Efstratiou. 2017. Next2Me: Capturing Social Interactions through Smartphone Devices using WiFi and Audio signals. In *Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (*MobiQuitous 2017*). Association for Computing Machinery, New York, NY, USA, 412–421. DOI:<https://doi.org/10.1145/3144457.3144500>
- [8] Geert Vanderhulst, Afra Mashhadi, Marzieh Dashti, and Fahim Kawsar. 2015. Detecting human encounters from WiFi radio signals. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia* (*MUM '15*). Association for Computing Machinery, New York, NY, USA, 97–108. DOI:<https://doi.org/10.1145/2836041.2836050>
- [9] Hande Hong, Chengwen Luo, and Mun Choon Chan. 2016. SocialProbe: Understanding Social Interaction Through Passive WiFi Monitoring. In *Proceedings*

of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services</i> (<i>MOBIQUITOUS 2016</i>). Association for Computing Machinery, New York, NY, USA, 94–103. DOI:https://doi.org/10.1145/2994374.2994387

[10] Jon Baker and Christos Efstratiou. 2017. Next2Me: Capturing Social Interactions through Smartphone Devices using WiFi and Audio signals. In <i>Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services</i> (<i>MobiQuitous 2017</i>). Association for Computing Machinery, New York, NY, USA, 412–421. DOI:https://doi.org/10.1145/3144457.3144500

[11] A. Mtibaa, K. A. Harras and M. Abdellatif, "Exploiting social information for dynamic tuning in cluster based WiFi localization," 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Abu Dhabi, 2015, pp. 868-875, doi: 10.1109/WiMOB.2015.7348053.

[12] Wang, Xueli & Kennedy-Phillips, Lance. (2013). Focusing on the Sophomores: Characteristics Associated With the Academic and Social Involvement of Second-Year College Students. *Journal of College Student Development*. 54. 541-548. 10.1353/csd.2013.0072.

[13] Mehrab, Zakaria & Ranga, Akhilandeshwari & Sarkar, Debarati & Venkatramanan, Srinivasan & Baek, Youngyun & Swarup, Samarth & Marathe, Madhav. (2020). High resolution proximity statistics as early warning for US universities reopening during COVID-19. 10.1101/2020.11.21.20236042.