1. **Protocol Description: Provide a brief description of the research in lay terms that can be understood by those unfamiliar with the area of research. (Richard)**

The objective of this project is to study large datasets of network logs and use it to identify patterns in people’s movements. Network logs are computer-generated records that can be used to tracks events and activities. Our goal is to better understand the way people move around campus (migration patterns), the way campus space is used (space utilization), and trends in internet usage in order to allow for better resource allocation and planning. Our approach to determine how people make use of the available spaces will be to look at the time and place in which they are logging in to a wireless network such as Wireless Network logs (WiFi), in conjunction with external data sources (e.g. demographics, weather, etc.)

This project consists of two stages. The first stage is finding the connection between WiFi data along with other external data and the number of people in an area (population inference). Once a heuristic is established, we will move to the second stage, which is developing and researching into applications of WiFi data inference, such as space utilization, WiFi usage trend, power consumption, etc. We will also be able to unveil patterns such as migration (how people move around) and clustering (where people tend to stay) habits for particular demographics (school affiliation, year, major, etc.) This will have a wide application to assist managerial decisions such as resource planning, network analysis, and sustainability.

1. **Describe the research design, including the proposed research methodology. For research directly involving human subjects, describe in chronological order the procedures that will occur. If subjects will be assigned to various conditions, describe how and why assignments will be made. (Examples of studies not directly involving human subjects, but still needing IRB approval, include prospective record reviews, observation of behavior without manipulation, and use of anonymized data). (Daniel)**

This study does not directly involve human subjects, but utilizes observation of space occupancy. There are two sources of data primarily. One dataset is supplied directly by the Georgia Tech Office of Information Technology (GT OIT) and includes network logs of user authentication times at various access points around campus. These data were anonymized with hashing functions prior to delivery to us and do not include any personally identifiable information. Secondary sources of information may include Infrared Proximity (IR) sensor information which is simply a binary indicator of whether a room is occupied, weather information, anonymized BuzzCard access times, and demographic information about the users such as year, major, etc. To reiterate, this information is tied to an anonymized user ID and cannot be directly linked to a user’s true identity.

The second primary source of data is the observations that we conducted in various venues around campus. These observations are time series counts of occupants of a room or open space (e.g. dining court or restaurant). We also collected the number of of open laptop devices. The objective was to determine how well correlated the observed counts were to the aforementioned WiFi data. Note that the subjects were not queried for information nor were they aware of the observations.

Once we have determined the relation between the ground truth (manual counts) and the data obtained through the WiFi network, we can build a predictive model to determine the amount of people that are projected to be at a certain time and place. This has numerous applications both theoretical and applied, some of which are outlined in the benefits section.

1. **State the duration of subject participation. How many hours, days, weeks or months? Specify number of sessions and, to the extent possible, state total amount of time for subject participation.**

This project will not require any direct human subject participation. The data is collected from log data and by observation, which will be conducted by the researchers. We expect to work with 3 years worth of WiFi network data. The observation will be conducted for a month, 3 hours per day.

1. **Describe study assessments and other data collection methods. Upload all instruments, including rating scales, questionnaires, surveys, focus group and interview guides, and so on at the end of this online application in the ATTACH DOCUMENTS section. (NOTE: The IRB recognizes that such specificity may not be possible in ethnographic or anthropological studies. In such cases, provide sufficient detail for the board to understand the study methodology). (Dan)**

As previously mentioned there are two primary sources of information: one supplied directly from the GT OIT which includes WiFi authentication, IR, and demographic information and the manual validation counts. The WiFi data may either record the authentication times and locations for anonymized users or the number of connected devices for a particular access point. This provides a further avenue for validation as these two sources can be compared. The objective of this study is to construct an accurate model for predicting occupancy counts from eclectic sources of information supplied to us by the GT OIT. Thus, the main assessment is to determine the accuracy of the model by comparing the predicted count to the observed count for various times and locations. This will help determine how well the model works under various conditions.

1. **Fully describe any potential benefits of this study. All ethical studies pose some benefit-- whether to individual research subjects, to the greater community, as a building block for further development of treatment, and so on. (If subjects will not benefit from participating, this should be disclosed in the benefits section of the consent document).(Dan)**

There are several potential benefits that could be derived from this study. Our objective is to construct an accurate model for the migration patterns on the Georgia Tech campus. This model could be used for a host of applications such as predicting space utilization at peak times, better automatic lighting and HVAC control, adaptive bus scheduling, or increasing power to overloaded wireless access points. These applications would have benefits for the Georgia Tech community as a whole.

In more detail, the model could be used to predict the occupancy (how many people or flow (how people move around) in study or dining areas at any given time. This would allow for better service planning and resource allocation such as study tables or checkout lines and could then be applied as a service (app, website, etc.) which would be useful intrinsically. It could also be used as a predictive or diagnostic tool for networks in general and serve as the basis for further studies.

1. **Fully describe any known risks to subjects participating in this study and, to the best of your knowledge, indicate the likelihood of such risks occurring. Also state any measures to be taken to minimize or eliminate risks or to manage unpreventable risks.(Richard)**

The data we are using in this research project come from two sources: 1) log data (WiFi authentication, BuzzCard transaction) that are recorded automatically through existing logging processes by the Office of Information Technology. 2) Passive observation at certain locations (e.g. Starbucks) recorded manually by the researchers. Therefore, the data collection part of the project does not pose any threats or consequences on human subjects. The personal information (such as user id, Mac-address of the devices) that are included in the log data will be de-identified by hashing functions before handed over to the researchers.

1. **Describe the statistical analysis plan, its design, and the rationale for the plan. (Daniel)**

Once we have all the required data, the statistical plan is to design a model (for example, a hidden Markov model or a conditional random field) to establish the relation between the observed data and the actual amount of people at a certain time and place. A classifier might also be needed to improve the accuracy once other publicly available data (such as the schedule of campus events, or the weather) is included. This could be later used as a predictive model to determine the amount of people at a certain location, as well as their affiliation to Georgia Tech.

1. **What are the anticipated start and end dates for the proposed research? Include the expected number of years that data analysis will continue. Federal regulations currently require that IRB approval remain active during data analysis (if subject data are not de-identified) even though subject enrollment and interaction may be complete. Be sure to include the period of data analysis when calculating the end date if you will maintain subject identifiers. (Richard)**

The project was started on May, 12th and it will last until July 28th. Human observation will be conducted for a month and the log data is expected to consist of 3 years worth of data. The analysis of the data will be expected to finish by the end of the year 2014.

Upload a fully annotated bibliography or reference section, including the results of the literature search done in support of this proposed study. This material may be added in the ATTACH DOCUMENTS SECTION at the end of the online application and is required for CLINICAL STUDIES only. (Daniel)

This is not a clinical study.

If this is a **student class project**, provide the course title and number and the name of the instructor.

**GEORGIA INSTITUTE OF TECHNOLOGY INVESTIGATORS ONLY:** If funding is pending, specify the potential funding source in the field here. (IRBWISE is linked only to active awards on record in the Georgia Tech Office of Sponsored Programs and not to pending proposals). If the study is already externally funded, please select the specific project in the funding section below.