**Data Management Plan**

**1. Data Type and Metadata Standard**

We will be using a combination of secondary use, de-identified, multi-sourced data combined with primary data collected via the randomized field experiments that we will design and implement. The work described in this proposal is expected to produce the following two broad categories of data: (i) Human Behavior Data and (ii) Computer Software (Source Code).

More specifically, the human behavior data from our research context include a combination of (1) individual demographics (e.g., age, gender, income level, education, home neighborhood); (2) individual job search activities as recorded by individual-specific logs and use of an online job-search tool, (3) employment status over time; (4) individual usage of ridesharing platforms (e.g., detailed trip data from Uber/Lyft); (5) individual detailed mobility GPS trajectory data (e.g., real-time geographic coordinates and time stamps recorded at regular intervals through a mobile application); (6) city map data and geo-spatial contextual data (e.g., type and density of points of interest). Note that, in some cases, particular data fields will need to be altered to preserve the privacy of experimental participants. For instance, GPS data that could be used to infer the home address of experimental participants will be randomly altered prior to dissemination in order to prevent the detection or deduction of individual identities. A complete IRB approval of our data disclosure plan will be obtained before the release of any data to the public.

The proposed work will also produce computer software and source code that we plan on making available. In particular, any experimental software that we will create to collect the data in this proposed research will be made available. Similarly, implementations of all the analytical algorithms will also be made available to the research community. Finally, any software independently developed for data integration, pre-processing, and visualization will also be made publicly available.

The raw data will be provided and documented in flat file format (e.g., CSV data files). We will integrate the multi-sourced data into common file formats available for downloading by the research community.

**2. Data De-identification and IRB Approval**

Due to the nature of the data in this proposed research, all individual-level data records will be de-identified before being transferred to the secured server for the researchers in our team to analyze them. We will not begin our experiment until full IRB approval for all aspects of our experimental study have been obtained. Events determined by the PIs to be unanticipated problems involving risks to subjects involved in the research, will be reported to the IRB per policy. If any protocol changes are needed, we will submit a modification request to the IRB. Protocol changes will not be implemented prior to IRB approval.

**3. Policies for Access and Sharing**

The PIs of this proposal are committed to widespread dissemination of the results and data. Within 2 months of the completion of the project, we will make all of our data (appropriately processed and deidentified in order to address privacy concerns) and tools available for other scholars who want to perform similar research or to replicate our work. We will make it a priority to make these data readily available and also to ensure that this fact is well advertised to the academic community.

We will publish the datasets and code with the agreement of the data providers. For those portions of data that are obtained from publicly available sources, no special precautions will be required to ensure security. For those datasets we obtain from proprietary sources, we will ensure the user information is completely de-identified. Moreover, we will apply appropriate privacy-preserving algorithms to ensure the data will not contain any sensitive or private information. The data will be freely available and no special authentication will be required for public access. Besides, we will publish the randomized user experimentation design and the original source code used to create any analytical estimation algorithms. We hope to help future researchers to adapt, modify and generalize the methods for their own use beyond the scope of this project. All associated documentation will be made available on a public project website.

We will utilize standard technologies and tools for database, analytics, and UI, and be compliant with industry standards.

**4. Plans for Archiving and Provisions for Appropriate Protection/Privacy**

We plan to allocate a robust, secure server infrastructure that can support the project site with minimal cost. The data server is expected to be 5 standard servers (Linux), with Intel (R) Xeon (R) Processor E7-4870 (30M Cache, 2.40 GHz, 6.40 GT/s Intel (R) QPI, 10 cores), 64 GB main memory and 10,000RPM server-level hard disks. MySQL will run on this server which provides secure data storage, good organization, data integration and analyses. Data will be regularly backed up by a DVD Data Disk to ensure continuity in the case of a catastrophic failure. We will maintain the site indefinitely, and the ongoing maintenance of the site infrastructure will not be tied to any single member of the research team. This will prevent changes in the roles of the project members from affecting the availability of the data or software.