Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. Rosetti, eds.

ON ELICITATION OF PREFERENCES FROM SOCIAL NETWORKS THROUGH SYNTHETIC POPULATION MODELLING

Przemysław Szufel
Bogumił Kamiński
Grzegorz Koloch

Warsaw School of Economics
Al. Niepodległości 162
02-554 Warszawa, Poland

ABSTRACT

Social network platforms are a useful source of information on preferences of citizens. However, population exposed on social network platform is non representative and in result preferences collected through such platforms are biased. The goal of the public administration is to utilize the data that can be collected through such online platforms in order to understand preferences and its structure in the society and hence better react to community's needs.

This situation calls for an algorithm that will allow to generalize information collected on social platform users on the entire population. We propose and evaluate a two-step methodology for testing of such algorithms: (1) synthetic population is generated and its sample is selected that represents social platform users and (2) regenerate the whole population on the basis of data from sub-population. In this way we can evaluate the quality of different algorithms aimed at preference elicitation from social platform data.

1 INTRODUCTION

Local governments are increasingly interested in improving communication with their citizens and with understanding their preferences. Therefore, they are implementing social collaboration platforms which allow members of their communities to discuss local governance issues. Such platforms allow for C2C (citizen to citizen) and C2G (citizen to government) communication. The goal of the public administration (PA) is to utilize the data that can be collected through such on-line platforms in order to understand preferences and its structure in the society and hence better react to community's needs. The challenge in development of algorithms for such situations is that usually PA lacks detailed information on general population. We propose a methodology for testing algorithms that allow to generalize information collected on social platform users on the entire population.

2 FORMAL MODEL

Consider an undirected network (graph) $G = (V,E)$. The ordered pair (V,E) comprises of a set of nodes (vertices) V and a set of edges E. Each node v, $v \in V$ represents a citizen in general population. An edge (v,u), $v,u \in V$ represents a social contact between two citizens. A network G contains k nodes, i.e. $|V| = k$. Let $N(v) = \{u : (v,u) \in E\}$ be a set of neighbors of node v. We assume that G is undirected, i.e. $(v,u) \in E \Leftrightarrow (u,v) \in E$ and does not have self-loops, i.e. $(v,v) \notin E$. For each citizen v, $v \in V$ we assign a socio-demographic data $\mathbf{d}(v)$, $\mathbf{d}(v) \in R^n$ and an opinion/preference $\mathbf{o}(v)$, $\mathbf{o}(v) \in \{0,1\}$.

Let $G^* = (V^*,E^*)$, $V^* \subset V$, $E^* \subset E$, $(v^*,u^*) \in E^* \implies u^*,v^* \in V^*$ be a sub-population of G (i.e. a group of users of a social web portal). We assume that for G^* we have information on connectivity

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

(E^*) that is revealed by contacts on the we portal, socio-demographic data $\mathbf{d}^*(v^*)$ and information on opinions/preferences $\mathbf{o}^*(v^*)$, $\mathbf{o}^*(v^*) \in \{0,1\}$.

The goal of the research is to identify a method to calculate the expected average preference value $\mathbf{o}(v)$ with regard to the demographic profile $\mathbf{d}(v)$. The available information includes sub-population web social network $G^* = (V^*, E^*)$, demographic data used to register $\mathbf{d}^*(v^*)$ and opinions revealed in the portal $\mathbf{o}^*(v^*)$. Moreover, we assume that a population census is available and hence distribution of the $\mathbf{d}(v)$ in the general population G is known.

3 ALGORITHM & SIMULATION RESULTS

A simulation procedure that enables testing different preference elicitation algorithms consists of the following meta-steps:

1. Generate a synthetic population that represents the "real population" for purposes of simulation experiments along with their social network, demographic profile and preferences; we assume that preferences of citizens depend on their intrinsic beliefs and preferences revealed by their neighbors in the network;
2. Select (not necessarily a representative sample) a sub-sample of the population that will serve as users of social platforms;
3. Apply algorithm reconstructing the starting population using data about the sample from step 2 and joint distribution of demographic profile of agents (we assume that it is known to PA from eg. census data);
4. Compare characteristics of the original synthetic population and reconstructed population.

We have performed a preliminary test of the above procedure using the following experiment setup.

1. Initial synthetic population was generated using the Watts-Strogatz algorithm; demographic profile consisted of two attributes: sex (with equal probabilities of male and female) and log-normally distributed income assuming that income depends on the node degree and sex of a citizen;
2. Opinion dynamic in the initial network was simulated assuming that agents have some initial preferences that can change conditional on their neighbors opinions;
3. A sub-population $G^* = (V^*, E^*)$ was randomly selected, with node selection probability proportional to its income (in this way we emulate sub-population non-representativeness);

As a test algorithm for reconstruction of the initial population we used the following approach. First we generated logistic regression model using the sub-population data to explain the probability of existence of links between the nodes conditional on their demographic profile. Next, we randomly generated the reconstructed population using known joint distribution of demographic profile and randomly created edges between citizens using the logistic regression. Finally we using data from the sub-population $G^* = (V^*, E^*)$ we build a model explaining dependence between citizen's opinion and her/his income, sex and social network status. The model is applied to calculate opinions for reconstructed population. Finally, we simulate the process of opinion propagation in the reconstructed network.

The preliminary results show that the proposed procedure allows not only to reconstruct population-wide opinion level but also to correctly identify the opinions in demographic sub-populations (in our example opinion conditional on citizen's sex and income).

ACKNOWLEDGEMENTS

This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 645860.