Computers in Human Behavior 51 (2015) 685–693

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Knowledge acquisition from social platforms based on network distributions fitting

Jarosław Jankowski [a,*], Radosław Michalski [b], Piotr Bródka [b], Przemysław Kazienko [b], Sonja Utz [c]

[a] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Żołnierska 49, 71-410 Szczecin, Poland
[b] Institute of Informatics, Wrocław University of Technology, Wrocław, Poland
[c] Knowledge Media Research Center, Tübingen, Germany

ARTICLE INFO

Article history:
Available online 27 December 2014

Keywords:
Social network analysis
Network sampling
Collaborative learning
Adaptive surveys

ABSTRACT

The uniqueness of online social networks makes it possible to implement new methods that increase the quality and effectiveness of research processes. While surveys are one of the most important tools for research, the representativeness of selected online samples is often a challenge and the results are hardly generalizable. An approach based on surveys with representativeness targeted at network measure distributions is proposed and analyzed in this paper. Its main goal is to focus not only on sample representativeness in terms of demographic attributes, but also to follow the measures distributions within main network. The approach presented has many application areas related to online research, sampling a network for the evaluation of collaborative learning processes, and candidate selection for training purposes with the ability to distribute information within a social network.

1. Introduction

Social networking sites are used as the research environment, and they provide opportunities to analyze real-world behavior (Abbasi, Chai, Liu, & Sagoo, 2012) as well as online activities (Gjoka, Kurant, Butts, & Markopoulou, 2009; Utz & Beukeboom, 2011) with the applications in the areas related to collaborative learning (Kwon, Liu, & Johnson, 2014), computer-mediated educational environments (Rummel & Spada, 2005) and knowledge management (Ordóñez de Pablos, 2004). Due to the complexity of the network structures, the analyses are usually performed using some samples to find structures that are smaller, but which share similar properties and distributions (Ebbes, Huang, Rangaswamy, & Thadakamalla, 2008). Recent studies in this field have focused on new algorithms (Lee, Kim, & Jeong, 2006; Stumpf, Wiuf, & May, 2005) and various areas of application (Gjoka et al., 2009, Lakhina et al., 2003, Rusmevichientong, Pennock, Lawrence, & Giles, 2001). The knowledge gathered from social network analysis can be extended using either typical surveys or new approaches based on adaptive surveys that optimize survey costs, quality and response rates. Research in this area is still in the early stages and adaptive methods are rarely implemented (Schouten, Calinescu, & Luiten, 2011). Another motivation for further research on the development of sampling methods is to increase the representativeness of survey data. The majority of studies on social media focuses on social network sites such as Facebook, and many of these studies use (online) surveys (Back, Stopfer, & Vazire, 2010; Utz & Krämer, 2009). The participants are usually students or self-selected. A problem with this approach is the representativeness of the sample – young, highly educated individuals or highly motivated users are usually overrepresented. Similar issues were identified in the field of knowledge management and collaborative learning to build groups with specific profile (Dascalua, Bodea, Lytras, Ordoñez de Pablos, & Burlacua, 2014). Although it is possible to extract behavioral data from social media and use them as the basis of the analysis (Liu, 2007; Thelwall, 2008), social scientists are often interested in the subjective experience of social media users, such as motivations for and gratifications of social media use, evaluation of competences and knowledge resources within the network (Colomo-Palacios, González-Carrasco, et al., 2014; Ordóñez de Pablos, 2004; Rózewski & Ciszczyk, 2009). To evaluate them, surveys are still the most suitable tool. In this paper, a new method for judging and enhancing the representativeness of an online sample is presented. The authors argue that it might be useful to utilize network measures such as centrality or degree as a basis for determining the representativeness of an online sample vs. the entire population.

* Corresponding author. Tel.: +48 91 449 56 68; fax: +48 91 487 08 42.
E-mail addresses: jjankowski@wi.zut.edu.pl (J. Jankowski), radoslaw.michalski@pwr.edu.pl (R. Michalski), piotr.brodka@pwr.edu.pl (P. Bródka), kazienko@pwr.edu.pl (P. Kazienko), s.utz@iwm-kmrc.de (S. Utz).

http://dx.doi.org/10.1016/j.chb.2014.12.015
0747-5632/

Some users have a very central social position within the online social networks, and they possess many more inbound and outbound connections when compared with other users. By comparing the network profile of the sample and the overall population, the representativeness of the online sample can be determined. Moreover, it is possible to develop algorithms that suggest which users should be approached in order to enhance the representativeness of a given sample so that the results will have higher potential in the areas of community building, information dissemination, and collaborative learning (Cowan & Jonard, 2004). The approach presented below is based on selecting an adequate set of candidates in each step of the multistage process to improve the representativeness of the sample in terms of network measures. Depending on the research goal and the area of applications, different network characteristics might be considered. To identify opinion leaders, the best candidates for leadership in collaborative learning or knowledge brokers, it is usually necessary to evaluate centrality measures (Boari & Riboldazzi, 2014). However, fulfilling a bridge position is more important when focusing on advertising and diffusing innovation or spreading knowledge among network nodes. From the perspective of collaborative learning, it is important to select nodes with specific characteristic for future activity within the network, and representative selection can impact on the future spread of knowledge within it.

While the structure of connections within the social network influences collaborative learning processes, there is a clear need to access information about participants and their potential for learning processes and sharing of information with other participants. Collaborative learning and group-based learning is closely related to dynamic social systems (Strijbos, 2001) where the members of the community interact and share experiences with one another (Chiu, 2008). During the learning process, members of the community evaluate other ideas and get engaged in monitoring the tasks and progress of other participants (Chiu, 2000). Key problems found here can be addressed to quantify proper users' features, select users with specific characteristic, and split users into optimal groups (Long & Qing-hong, 2014) in order to boost the sharing of knowledge in organizations (Lytras, Tennyson, & Ordóñez de Pablos, 2008). During collaborative learning processes, building teams and increasing potential by acquiring additional representatives with specific knowledge or competences can be very important, not only in terms of knowledge itself, but also in terms of network characteristics. While the ability to attain knowledge from all nodes of a network can be limited, sampling methods can be applied to acquire information desired. The proposed method can be adapted to different research goals by using weighted sampling. As online surveys are usually based on voluntary participation, and because there may be low response rates, the obtained sample may have other characteristics than the random sample. The proposed method makes it possible to direct the selection process towards expected characteristics of the sample.

2. Related work

2.1. Conventional and adaptive network sampling

Research related to network sampling is based on various techniques using both conventional and adaptive approaches. Sampling design is treated as *conventional* when it does not use acquired data in the sampling process. The first group of methods in this class is based on random-node selection focused on uniform or proportional-to-node degree probabilities (Maiya & Berger-Wolf, 2010), random edge selection (Ahmed, Neville, & Kompella, 2011) and the egocentric method (Ma, Gustafson, Moitra, & Bracewell, 2010). The other group is based on graph sampling and includes snowball sampling (Frank, 1979; Frank & Snijders, 1994) random walk (Thompson, 1998), the forest fire method (Leskovec & Faloutsos, 2006) and others. Apart from theoretical work, some studies were conducted using real online social systems like Facebook (Gjoka et al., 2009) or Twitter (Ahn et al., 2007).

In contrast to static designs, *adaptive* sampling can be applied after the results of earlier stages are collected, and it is used to direct sampling (Handcock & Gile, 2010; Thompson, 2011). Conventional methods have problems with sampling hidden populations, but the adaptive method can change sampling direction on the fly, if necessary. There are approaches targeted to adaptive cluster sampling based on the selection of neighbors in the network only if a given condition related to cluster location is satisfied (Thompson, 1998). Other dedicated methods are used to sample network node selection and the estimation of information diffusion processes in either single-layer (Jankowski, Michalski, & Kazienko, 2012) or multilayer networks (Michalski, Kazienko, & Jankowski, 2013).

The respondent-driven sampling was introduced by Heckathorn (2007) and extended later (Salganik & Heckathorn, 2004). It is based on recruitment of members of the population by other sampled members. Respondent-driven sampling is an extension of snowball sampling and the patterns of recruitment are used to calculate inclusion probabilities for different types of nodes. It collects information about ties from each participant, but can be inaccurate in clustered networks because of homophily and separated communities. The proposed adaptive approach is based on the collection of network data from respondents, and adaptive sampling (Thompson & Seber, 1996) is based on moving to other regions of the network after obtaining enough samples from the identified cluster.

2.2. Adaptive approaches to survey design

While sampling delivers information about the network evolution of data collection methods, new technologies provide possibilities for survey design that were unavailable earlier (Deville & Tillé, 2005). Surveys can be identified as *static* if they are not dependent on collected observations, while *adaptive* surveys are partially based on data from observations (Schouten et al., 2011). Adaptive surveys are a means of increasing responses and the quality of the research by selecting samples characterized by the lowest mean square error on the sample values. Apart from sampling direction, other adaptive components can be: offering different incentives, using responsive survey designs (Groves & Heeringa, 2006) or questionnaire structures (Singh, Howell, & Rhoads, 1990). Survey adaptation can be based on time intervals between calls, visits and other forms of communication with respondents (Greenberg & Stokes, 1990), survey errors (Lyberg et al., 1997) and survey costs (Groves, 1989). The design-based approach to survey sampling uses variables of interest as fixed values, while model-based variables of interest are defined as random variables with joint distribution (Thompson, 1998). During surveys, interventions can be made to decrease variances of selected variables in the respondent pool by targeting sampling to key subgroups (Couper & Groves, 2009).

Earlier research showed how to optimize the survey process and increase response rates (Schouten et al., 2011). Schouten et al. systematized adaptive survey designs and provided a mathematical framework to improve the process of data collection based on surveys. Furthermore, the authors defined $Q(p)$ as an indicator of quality and $C(p)$ as a cost indicator, and optimization was defined as $max_p\ Q(p)$ with $C(p) < Cmax$ and $Cmax$ as maximum budget constraints or $min_p\ C(p)$ with $Q(p) \geqslant Qmin$ and $Qmin$ defined as minimum quality. The quality functions can be