Sampling Bias in LinkedIn: A Case Study

Shanshan Zhang
Temple University
1805 N Broad Street
Philadelphia, USA
shanshan.zhang@temple.edu

Slobodan Vucetic
Temple University
1805 N Broad Street
Philadelphia, USA
vucetic@temple.edu

ABSTRACT

This paper describes a case study of sampling bias in LinkedIn, a major professional social network. The study collected a sample of 1,989 STEM students who graduated from a major public university between 2002 and 2014. Overall, 40% of the graduates had a LinkedIn profile in summer of 2015. It was observed that LinkedIn participation significantly fluctuated among different majors, and ranged from 30% for biochemistry majors to 51% for information science majors. Year of graduation, gender, and grade point average surprisingly did not seem to create a large difference in LinkedIn participation. These results should be useful for design and interpretation of empirical studies which use LinkedIn data or select participants from LinkedIn social network.

Categories and Subject Descriptors

J.4 [SOCIAL AND BEHAVIORAL SCIENCES]: Sociology

Keywords

LinkedIn; sampling bias; major; year; gender; GPA

1 Introduction

There is an increasing interest in using rich information from popular online social networks to perform quantitative and qualitative empirical research in social sciences. There have been over 400 active research studies of the Facebook ecosystem [4] as well as numerous studies using other popular social networks such as Twitter and LinkedIn. Those empirical studies have been taking advantage of the unprecedented sample sizes, richness of demographic and behavioral information, and ease of access to study participants. Among the published survey-based studies using LinkedIn, the largest professional social network, we mention [2], in which authors posted a questionnaire in LinkedIn groups to investigate the habits of code example usage of professional programmers, and [3], in which LinkedIn groups were used to define the role of software architect. An example of large-scale studies using LinkedIn profiles is [6], which analyzes career trajectories.

While easy access to study participants or to large quantities of data is tempting, it is very important to understand possible sources of sampling bias when designing the empirical studies and interpreting their results. As pointed out in [1], social network usage is not uniformly distributed among the population, with differences depending on gender, age, and role (individual contributor vs. manager). The objective of this paper is to study sampling bias in LinkedIn, from the perspective of recent bachelor's degree holders from STEM fields. We were interested in learning whether sampling bias exists with respect to major, year of graduation, gender, and grade point average. The sample for this case study were recent graduates from a major public university.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889357.

2 Data Set

Subjects. We considered bachelor's degree holders who graduated from a major public university between 2002 and 2014 and who majored in one of 7 STEM fields. For each of the 13 years of graduation and the 7 majors, we randomly selected up to 30 graduates. This resulted in the total sample size of 1,989 graduates.

Attributes. For each graduate, we collected 3 categorical and 1 numerical attribute:

- Gender: Female or Male.
- Graduation Year: between 2002 and 2014.
- Major: Biology (BIO), Computer Science (CS), Biochemistry (BIOCH), Information of Science and Technology (IST), Chemistry (CHEM), Geology (GEO), and Mathematics (MATH).
- Cumulative Grade Point Average (GPA).

LinkedIn label. For each of the 1,989 graduates, we determined if they had a LinkedIn profile during summer of 2015. To determine this, we manually entered each graduate's name together with words LinkedIn and [university name] into Google search. For example, if the name was "John Doe", our entry was "John Doe [university name] LinkedIn". If there were one or more LinkedIn matches among the top 20 results, we checked if the public LinkedIn profile matched the university name, year of graduation, and major of the graduate. If no matches were found and the graduate was female, we repeated the search by removing the last name, to account for the possibility that the graduate changed the last name.

3 Sampling Bias Analysis

Major Bias. In Figure 1 we show the fraction of LinkedIn users among graduates from each of the 7 majors, where the number on top of each bar is the sample size. We can see that there is a sizeable difference in the fractions; bio-related majors have the lowest participation (around 30%), while

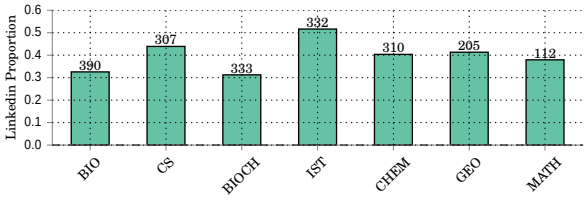

Figure 1: Participation rate by major

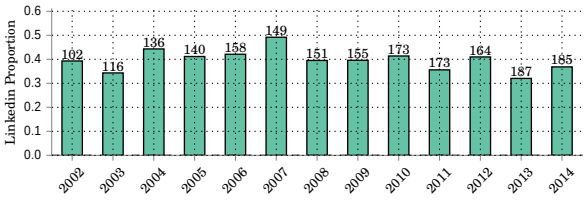

Figure 2: Participation rate by graduation year

computer and information science majors have the largest participation (over 45%) in LinkedIn. For the whole sample of 1,989 graduates, the LinkedIn participation rate is 40%. To calculate the statistical significance, in Table 1 we show the p-values of the pairwise comparisons between majors. For each pair of majors m_1 and m_2, we performed *one-tail randomization test for two proportions* [5]. The reported p-value is the fraction of times the difference between the two majors is smaller than the difference between the majors where LinkedIn label is randomly permuted. It could be seen that many of the pairs exhibit significantly different LinkedIn participation at 0.05 significance level.

Graduation Year Bias. In Figure 2 we show the fraction

Table 1: Pair-wise comparison of majors

	BIO	CS	BIOCH	IST	CHEM	GEO
CS	**0.01***	-	-	-	-	-
BIOCH	0.37	**0.001****	-	-	-	-
IST	**0.00****	**0.02***	**0.00****	-	-	-
CHEM	**0.02***	0.19	**0.01***	**0.003***	-	-
GEO	**0.02***	0.27	**0.02***	**0.01***	0.43	-
MATH	0.17	0.16	0.13	**0.01***	0.35	0.32

of LinkedIn users for each graduation year between 2002 and 2014. We can observe minor differences in the participation rates that fluctuate around 40%. To measure significance we define $P_{\leq y_0}$ as the participation rate of students that graduated before year y_0 and $P_{>y_0}$ after y_0. Using the randomization test, only 2010 had p-value for 2010 below 0.05, indicating a potential slight drop in LinkedIn participation of students that graduated after 2010.

Gender Bias. In Table 3 we compare the LinkedIn par-

Table 2: Testing graduation year bias

y_0	$P_{\leq y_0}$	$P_{>y_0}$	p	y_0	$P_{\leq y_0}$	$P_{>y_0}$	p
2002	0.37	0.39	0.29	**2008**	0.41	0.38	0.08
2003	0.37	0.40	0.16	**2009**	0.41	0.37	0.05
2004	0.39	0.39	0.46	**2010**	0.41	0.37	**0.04***
2005	0.39	0.39	0.42	**2011**	0.40	0.37	0.14
2006	0.40	0.39	0.21	**2012**	0.40	0.36	0.05
2007	0.41	0.38	0.10	**2013**	0.39	0.37	0.24

ticipation rates of females and males in each of the 7 STEM majors. As can be seen, for most majors there is no significant difference (using the previously described randomization test) between females and males. The only significant difference is in BIO (with larger participation of females) and IST (with larger participation of males). Overall, our conclusion is that there is no obvious within-major gender bias in LinkedIn.

GPA Bias. In Table 4 we compare average GPA of

Table 3: Testing gender bias

	$Count_M$	$P_M(m)$	$Count_F$	$P_F(m)$	p
BIO	134	0.26	256	0.36	**0.03***
CS	271	0.44	36	0.36	0.20
BIOCH	151	0.30	182	0.33	0.30
IST	258	0.55	74	0.39	**0.005****
CHEM	139	0.35	171	0.41	0.16
GEO	92	0.38	113	0.43	0.20
MATH	77	0.34	35	0.40	0.25
OVERALL	1122	0.40	867	0.38	0.10

graduates in each major depending on their LinkedIn participation. Instead of reporting the actual GPA, to preserve data privacy, we converted GPAs of graduates in each (year, major) group to percentiles. As can be seen, LinkedIn members from BIOCH and IST majors seem to have a slight statistically significant increase in GPA over the non-LinkedIn members. LinkedIn members from all majors but BIO seem to have slightly larger average GPA than non-LinkedIn members. However, the difference on the overall population of 1,989 graduates was not significant at 0.05 significance level.

Conclusion. Our study reveals that there is a significant

Table 4: Testing GPA bias

m	$Y_{NL}(m) \pm std$	$Y_L(m) \pm std$	p
BIO	51.3 ± 1.8	47.8 ± 2.4	0.14
CS	49.4 ± 2.1	51.2 ± 2.6	0.30
BIOCH	48.4 ± 2.0	54.0 ± 2.5	**0.037***
IST	47.1 ± 2.4	52.9 ± 2.1	**0.039***
CHEM	49.2 ± 2.1	51.8 ± 2.6	0.24
GEO	48.5 ± 2.6	52.8 ± 3.2	0.17
MATH	50.4 ± 3.4	50.5 ± 4.6	0.48
OVERALL	49.2 ± 0.8	51.3 ± 1.0	0.07

difference in LinkedIn participation among different STEM majors, while differences based on year of graduation, gender, and GPA are minor or insignificant.

References

[1] A. Archambault and J. Grudin. A longitudinal study of facebook, linkedin, & twitter use. In *SIGCHI*, 2012.

[2] O. Barzilay, O. Hazzan, and A. Yehudai. Using social media to study the diversity of example usage among professional developers. In *SIGSOFT*, 2011.

[3] N. Unkelos-Shpigel, S. Sherman, and I. Hadar. Finding the missing link to industry: Linkedin professional groups as facilitators of empirical research. In *International Workshop on CESI*, 2015.

[4] R. E. Wilson, S. D. Gosling, and L. T. Graham. A review of facebook research in the social sciences. *Perspectives on psychological science*, 7(3):203–220, 2012.

[5] T. H. Wonnacott and R. J. Wonnacott. *Introductory statistics*, volume 19690. Wiley New York, 1972.

[6] Y. Xu, Z. Li, A. Gupta, A. Bugdayci, and A. Bhasin. Modeling professional similarity by mining professional career trajectories. In *KDD*, 2014.