RESEARCH

Collective Aspects of Privacy in the Twitter Social Network - Supporting Information

David Garcia*, Mansi Goel, Amod Agarwal and Ponnurangam Kumaraguru

*Correspondence: garcia@csh.ac.at
Full list of author information is available at the end of the article

Principal Component Analysis

Our purpose with the application of PCA to the 100-dimensional Doc2Vec results is to reduce the dimensionality of the feature space to have a denser representation. Figure 1 reports the percentage of variance explained by each Principal Component. We selected the number of Principal Components by searching for the elbow of the curve on the differences between Principal Component variances. After the second most explanatory Principal Component, the variance drops are below 1%, more than an order of magnitude below the variance explained by the most explanatory Principal Component (12%). Note that we do not claim that this representation is highly informative of user biographical data, we only perform this step to simplify the analysis of the shadow profile hypothesis for biographical texts. Models with higher dimensionality of the Doc2Vec representation and more Principal Components beyond two can be more informative in future studies with larger samples of ego users.

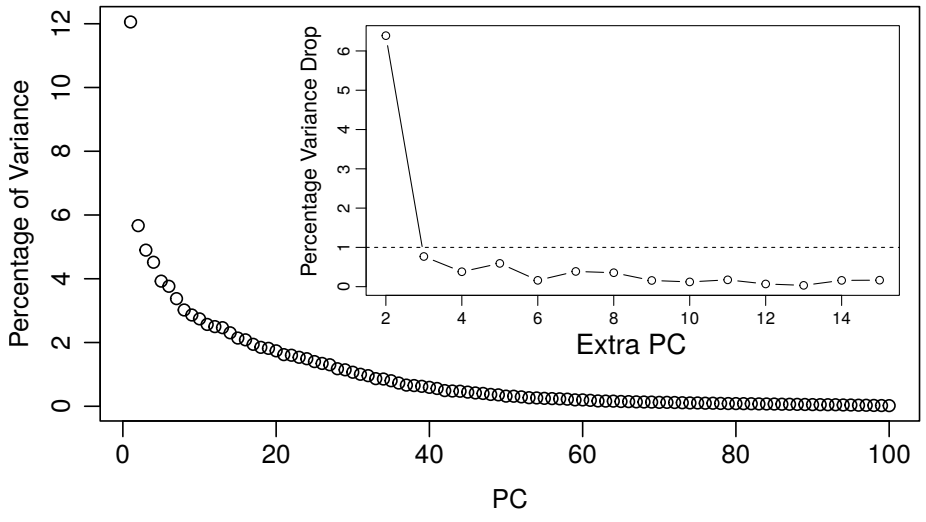

Figure 1 Percentage of variance explained by each Principal Component (PC). The inset shows the difference between consecutive Principal Components. We set a threshold of 1% variance on the differences, selecting this way the first 2 Principal Components.

Regression Procedure

All regression models were formulated as Bayesian regression models in the bayesglm function of the arm R package [1] with weakly informative prior distributions [2] for all parameters. 95% Credible Intervals were computed over 10000 simulations after convergence and R^2 values were computed as the fraction of explained variance of the dependent variable.

Regression errors and amounts of alters were log-transformed to reduce their skewness. For the analysis of disclosure tendencies, predictor and Null Model errors and cosine similarities were calculated over 1000 samples of disclosing users and 1000 randomizations of the Null Model.

Regression Results

term	estimate	95% CI
Intercept	$\mathbf{5.44}^{***}$	$[4.09, 6.76]$
$log(N_{alters})$	$\mathbf{-0.49}^{*}$	$[-0.90, -0.06]$
AIC	3615.2	
BIC	3628.3	
R^2	0.009	
N	584	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 1 Regression results for the logarithm of empirical error of shadow profiles of location as a function of the logarithm of the amount of disclosing alters of each ego user ($log(N_{alters})$).

	N_{alters} Model		Model with interaction	
term	estimate	95% CI	estimate	95% CI
Intercept	$\mathbf{5.57}^{***}$	$[5.1, 6.04]$	$\mathbf{5.5736}^{***}$	$[5.1, 6.05]$
$log(N_{alters})$	$\mathbf{-0.48}^{***}$	$[-0.62, -0.34]$	$\mathbf{-0.35}^{***}$	$[-0.52, -0.19]$
$\rho * log(N_{alters})$			$\mathbf{-0.24}^{**}$	$[-0.39, -0.09]$
AIC	37570		37562.6	
BIC	37590		37589.3	
R^2	0.008		0.009	
N	5923		5923	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2 Regression results for the models of the logarithm of errors as a function of $log(N_{alters})$ and its interaction with the disclosure parameter ρ.

	N_{alters} Model		Model with interaction	
term	estimate	95% CI	estimate	95% CI
Intercept	-0.03	$[-0.08, 0.01]$	-0.03	$[-0.08, 0.01]$
$log(N_{alters})$	$\mathbf{0.028}^{***}$	$[0.01, 0.04]$	$\mathbf{0.021}^{*}$	$[0.01, 0.04]$
$\rho * log(N_{alters})$			0.011	$[-0.004, 0.027]$
AIC	12246.4		12246.4	
BIC	12266.7		12273.5	
R^2	0.002		0.002	
Num. obs.	6470		6470	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3 Regression results for the models of cosine similarity of prediction and biographies as a function of $log(N_{alters})$ and its interaction with the disclosure parameter ρ.

Author details

References

1. Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M.G., Kerman, J., Zheng, T., Dorie, V., Su, M.Y.-S.: Package 'arm' (2016)
2. Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S.: A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 1360–1383 (2008)