Appendix

Keywords
This subsection lists the search strings that we used to identify posts in each of the topics conflict, protest, strike, corruption and political positions/politicians.

Table A1: Conflict, protests and strikes

Conflict	Protest	Strike
被袭击	堵路	罢弛
被袭击 and (政府 or 官员 or 干部)	非法集会	罢工
威胁政府	集会 and (群众 or 公众 or 大规模)	罢课
催泪弹 and (群众 or 政府 or 警察)	静坐	罢驶
官民 and (矛盾 or 冲突 or 暴力 or 对抗)	请愿	罢市
集体冲突	示威	罢运
警民 and (矛盾 or 冲突 or 暴力 or 对抗)	讨薪	
军民 and (矛盾 or 冲突 or 暴力 or 对抗)	学潮	
镇压	工潮	
	游行	
	学生 and 闹事	
	封堵 and (政府 or 群众 or 工人 or 公路)	
	自焚	
	千人下跪	
	not 反日	
	not 抗日	
	not 反日	

Table A2: Corruption

Corruption
腐败 and (政府 or 部门 or 官员 or 干部 or 官员)
腐败分子
公款
贿赂
廉政
买官
卖官
挪用
社保 and (贪污 or 腐败 or 挪用)
受贿
索贿
贪污
行政腐败
徇私
滥用职权
利益集团
侵占 and (政府 or 官员 or 部门 or 干部)
情妇 and (政府 or 官员 or 部门 or 干部)
失职 and (政府 or 官员 or 部门 or 干部)
私分 and (政府 or 官员 or 部门 or 干部)
私生 and (政府 or 官员 or 部门 or 干部)
伪造 and (政府 or 官员 or 部门 or 干部)
舞弊 and (政府 or 官员 or 部门 or 干部)
虚报 and (政府 or 官员 or 部门 or 干部)
虚开 and (政府 or 官员 or 部门 or 干部)
诈骗犯 and (政府 or 部门 or 官员 or 干部)
诈骗罪 and (政府 or 部门 or 官员 or 干部)

Table A3: Politicians

Political position/person	Keywords
Xi Jinping	习近平
Xi Jinping	习大大
Xi Jinping	习总
Li Keqiang	李克强
Hu Jintao	胡锦涛
Wen Jiabao	温家宝
Wen Jiabao	温总理
Provincial governor	省长 or 区主席 or 省主席 or 区副主席 or 省副主席
Provincial Party secretary	(书记 and (省委 or 自治区)) or 省书记 or 省副书记
City mayor	市长 or 州主席 or 州专员 or 地区专员
City party secretary	(书记 and (市委 or 地委 or 自治州)) or 市书记 or 市副书记
County governor	县长
County Party secretary	书记 and 县委
Village chief	村长
Village Party secretary	村支书

Identifying government affiliation by language

We identify 1,042 official, government-affiliated and 538 newspaper accounts by manually inspecting the blogs of thousands of users with user names typically associated with these functions. We then estimate a Support Vector Machine (SVM) to identify these users from a 1 percent sample (28,440) of randomly drawn users based on frequencies of certain words in their posts.

The word frequencies in each post are computed after the pre-processing described at the earlier section in note 4. Based on performance in other classification tasks, SVMs have been identified as one of the most efficient classification methods (Dumais et al., 1998, Joachims, 1998, Sebastiani, 2002). As inputs to the SVM, we use term-frequency inverse document frequencies. We use the software SVM-light Joachims (1999). In the SVM classification, a large number of words are important. However, just to give a sense of the classification, the words with the highest weight are "Communist Youth League", "Municipal Party Committee" and "Convention". To assess how well the SVM performs we use cross-validation where we leave iteratively leave out 1 government account and 17 non-government accounts, estimate the model and classify the left-out observations. This classifier has a precision of .81 and a recall of .41. A more familiar statistic is perhaps the t-statistic of a probit regression of a variable indicating a dummy account on the SVM-output parameter used for classification. This t-statistic is 56, meaning that language is highly predictive of government accounts.

Since government accounts were over-sampled in the above estimation sample, we cannot use it to estimate the share of government accounts. We instead draw a new random sample of 500 users. In this sample, we estimate a probit model of the probability of being a government account conditional on the SVM parameter; see Table A4 below. This process is known as Platt scaling and is a common way to map the SVM parameter estimates to probabilities (Platt, 1999). We combine the SVM-parameters with the probit estimates to predict the probability that each account is a government account.

Table A5. Dependent variable: government account dummy

VARIABLES	I
SVM	3.548***
	(0.752)
Constant	1.799**
	(0.768)
Observations	500

Unit of observation is a Weibo account. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1