Political Analysis (2015) 23:299–305
doi:10.1093/pan/mpv005

Estimating the Severity of the WikiLeaks U.S. Diplomatic Cables Disclosure

Michael Gill
Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA, 02138
e-mail: mzgill@fas.harvard.edu (corresponding author)

Arthur Spirling
Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA, 02138
e-mail: aspirling@gov.harvard.edu

Edited by Jens Hainmueller

In November 2010, the WikiLeaks organization began the release of over 250,000 diplomatic cables sent by U.S. embassies to the U.S. State Department, uploaded to its website by (then) Private Manning, an intelligence analyst with the U.S. Army. This leak was widely condemned, including by then Secretary of State Hillary Clinton. We assess the severity of the leak by considering the size of the disclosure relative to all diplomatic cables that were in existence at the time—a quantity that is not known outside official sources. We rely on the fact that the cables that were leaked are internally indexed in such a way that they may be treated as a sample from a discrete uniform distribution with unknown maximum; this is a version of the well-known "German Tank Problem." We consider three estimators that rely on discrete uniformity—maximum likelihood, Bayesian, and frequentist unbiased minimum variance—and demonstrate that the results are very similar in all cases. To supplement these estimators, we employ a regression-based procedure that incorporates the timing of cables' release in addition to their observed serial numbers. We estimate that, overall, approximately 5% of all cables from this timeframe were leaked, but that this number varies considerably at the embassy-year level. Our work provides a useful characterization of the sample of documents available to international relations scholars interested in testing theories of "private information," while helping inform the public debate surrounding Manning's trial and 35-year prison sentence.

1 Motivation

For scholars of international relations, the WikiLeaks U.S. diplomatic cable disclosure of 2010 is a promising source of data on the American government's beliefs and strategies as they pertain to interactions with other states. A mainstay of the now "dominant" (Lake 2010), "rationalist" approach to war (Fearon 1995), this type of "private information" is by its very nature difficult to obtain for analysts: it may never be released, on the grounds that to do so would damage national interests, endanger its citizens (though see Shapiro and Siegel 2010 for discussion), or more cynically, because it might allow popular challenge to elite control of foreign policy (Gibbs 1995). While this has not necessarily prevented data-driven scholarship on the canonical bargaining model of conflict (e.g., Fearon 1994b; Werner 1999; Reed 2003; Smith and Stam 2004; Ramsay 2008; Reiter 2009) or historical case studies of important incidences of decision-making (e.g., Goemens 2000; Snyder and Borghard 2011), it does mean that such work cannot directly examine all the incentives faced, and actions taken, by leaders in the contemporary period. Precisely because it contains documents that were not designed for public release and thus are surely more candid about particular political relationships, the WikiLeaks disclosure potentially represents a trove of data that might be tapped to test subtle theories in international relations.

Authors' note: Jonathan Bennett and Michael Egesdal provided helpful comments on an earlier draft. The authors are grateful to the Office of the General Counsel at Harvard University for legal advice. Replication materials (code, not data) for the estimators they describe can be found in the *Political Analysis* DVN archive: see Gill and Spirling (2015).

© The Author 2015. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mpv005 Published online by Cambridge University Press

These include hypotheses regarding "audience costs" (Fearon 1994a) and the signaling of resolve (see e.g., Smith 1998; Schultz 1999; Slantchev 2006).

As with any "new" data source, it is helpful to understand the nature of the sample available for researchers before they conduct statistical analysis. A fundamental question in this respect is the size—i.e., the "severity"—of the leak (sample) relative to all cables available (the population) at the time. If this sample is "small"—and far from the universe of cases—then some caution is required in interpreting results using the documents. This is especially true when examining particular embassies (in particular countries) in particular years for which the data are necessarily more finely sliced. For example, if researchers have an interest in recent U.S. diplomatic policy in certain areas of the world (e.g., Christensen 2006), the utility of studying the disclosed cables presumably varies according to the density of their coverage: here, knowing that one has access to 5% of all instructions given to ambassadors is very different to having 95% or 100% of those orders. Furthermore, knowledge of population size is helpful since it prompts further investigation of the sampling process itself: that is, it encourages analysts to think carefully about possible selection biases—especially for data such as these that were not systematically released according to well-known bureaucratic rules—and adjust for them in their statistical approaches.

Quite separate to these methodological concerns, the Manning case received widespread public attention and broader interest, including accusations of inhumane treatment from the United Nations special rapporteur on torture.[1] Whether or not Manning's sentence was just depends in part on the purported severity of the crimes she committed, and in part on the normatively appropriate relationship between the state and secrecy—a topic of recent interest to political theorists (see Sagar 2013). Since this Letter may be seen as an attempt to measure the first of these issues, we see it as a contribution to the public debate surrounding the trial.

2 Data

The WikiLeaks data consist of 251,237 diplomatic cables sent by the U.S. State Department to U.S. embassies and missions around the world. Machine-readable versions of these documents are available at various websites, but what is *not* known to researchers is the fraction of all cables in existence that the leak represents. The original data date range is from 1966 to 2010, but we focus on all cables released on or after January 1, 2005. We do this because coverage prior to the year 2005 is sparse (there are approximately as many cables from 2005 to 2010 as there from 1966 to 2005); second, because we wished to guard against any change in protocols—concerning the nature of the cables—that the terrorist attacks of September 11, 2001, may have ushered in and that we assume were in full force by 2005. This leaves 163,958 cables.[2]

Cables are classified into three categories, depending on the degree of damage to national security that "the unauthorized disclosure of which reasonably could be expected to cause."[3] In descending order of the purported balefulness of unauthorized release, these categories are "Top Secret,"[4] "Secret," and "Confidential." Any cable not meeting the criteria for such restricted access is "Unclassified."

From inspection we note that the cable titles, known internally as the "Reference ID," follow a consistent naming convention. Examples of this pattern can be seen in Table 1: a sample from the U.S. Embassy in Tokyo for 2009. Consider the Reference ID `09TOKYO13`: the first two digits connote the year of creation (2009) (confirmed by the date "Created" column—information that we culled from the original cable texts); this is followed by the issuing embassy (`TOKYO`), followed by a number (`13`). This particular cable was unclassified, as can be seen in the third column. Our immediate concern is whether the "`13`" demarcates that this was, indeed, the 13th document sent by the Embassy that year, and thus whether we can treat all such numbers as indicative of their position in the creation order. Our primary check on this logic was to plot the document numbers

[1] "Bradley Manning's treatment was cruel and inhuman, UN torture chief rules," *The Guardian*, March 12, 2012.
[2] Our replication files are as described in Gill and Spirling (2015).
[3] Executive order 13526, 2009.
[4] None of our data have this classification.

https://doi.org/10.1093/pan/mpv005 Published online by Cambridge University Press

Table 1 Example of reference ID naming convention: two-digit year, followed by embassy origin, followed by cable number

Reference. ID	*Date created*	*Classification*
09TOKYO4	January 2, 2009	Confidential
09TOKYO8	January 5, 2009	Unclassified
09TOKYO12	January 5, 2009	Unclassified
09TOKYO13	January 5, 2009	Unclassified
09TOKYO15	January 6, 2009	Unclassified
09TOKYO17	January 6, 2009	Unclassified
09TOKYO20	January 6, 2009	Unclassified

against the dates of issue (Column 2) for a sample of embassies, and verify that higher numbers appear on later dates. With some rare exceptions, this is the case.[5] Where it is not the case, the reason is likely one or more of the following: multiple cables are sent from the same embassy on the same day; cables of different secrecy statuses are processed or cleared at different rates; it takes longer for some cables to be started and completed than others; or there is some mild seasonality in the rate at which embassies write cables. We do not know why some cables are missing from the leak; for example, the sample implied by Table 1 is $\{4, 8, 12, 13, 15, 17, 20\}$, implying that cables with numbers 5, 6, 7, 9, 10, etc., were written but not leaked. Two possibilities seem plausible: Manning had limited access to the full population of diplomatic communications, whereby the "unobserved" cables were unavailable at her level of security clearance (e.g., this includes those classified as "top secret"); alternatively, Manning was not systematic in gathering the cables, and thus we receive a *de facto* random sample of those available.

If the cable numbers are a random sample from a discrete uniform distribution (an assumption we assess below), then there are well-known techniques for estimating the total number of such cables in existence at the time of disclosure. The next section explains how.

3 Estimating the Maximum of a Discrete Uniform Distribution: The German Tank Problem

Popular in introductory statistical texts (e.g., Larsen and Marx 2011), the German Tank Problem considers the position of a commander of Allied Forces in the Second World War seeking to estimate the total number of enemy tanks (N) in existence (see Ruggles and Brodie 1947 for a detailed historical discussion). The commander has observed the serial numbers of a (without replacement) random sample of such tanks, of size k, and has only this information on which to base his estimate. It may be assumed that the serial numbers are sequentially numbered in the population, starting with one for the first tank produced by the enemy, two for the second tank, and so on. Of course, the commander does not necessarily observe any consecutive numbers in his sample and a typical (for $k = 6$) sample might be as follows: $x = \{4, 30, 14, 5, 55, 41\}$.

Generally, the problem is to estimate the maximum of a discrete uniform distribution using sampling without replacement. The maximum-likelihood estimate, $\hat{N}_{\mathrm{MLE}}$, is well known, and is simply the maximum of the sample itself, denoted m. So, $\hat{N}_{\mathrm{MLE}} = 55$ tanks for our running example. Derivations may be found elsewhere, but the intuition is straightforward: clearly, the maximum-likelihood estimate cannot be less than the maximum of the sample, since this implies that at least one of the sample values is impossible. More subtly, $\hat{N}_{\mathrm{MLE}}$ also cannot be a number larger than the maximum of the sample because that number was not observed, and thus plugging it

[5] We also verified for each that the ID number for a given embassy resets to 0 at the start of a new year (e.g., moving from 2008 to 2009 for Tokyo cables).

https://doi.org/10.1093/pan/mpv005 Published online by Cambridge University Press

into the likelihood must return a lower value than would be seen if m was used.[6] We note that m is also a sufficient statistic for N.

While $\hat{N}_{\text{MLE}}$ is straightforward to calculate, and consistent, it is biased. Formal proofs are not complicated, but the intuition suffices for current purposes: because m is always less than or equal to the true value of N, the sample maximum will not, in small samples, reveal the truth "on average." Goodman (1952) suggests a minimum variance unbiased estimator $\hat{N}_{\text{Goodman}} = m + (m/k) - 1$. The key component, m/k, represents the "average gap between ... the serial numbers" (Goodman 1952, 622–23), and adjusts the maximum-likelihood estimate upward in a way that is consistent with the information in the sample. Goodman (1952) then gives an unbiased variance estimator; formulae for confidence intervals are given in Goodman (1954). For our running example, $\hat{N}_{\text{Goodman}} \approx 63$.

Höhle and Held (2006) give a Bayesian estimator, wherein the goal is to obtain a posterior distribution, Pr $(N|m)$. From Bayes's rule, we have that Pr $(N|m) =$ Pr $(m|N)$Pr (N)/Pr (x), where Pr (x) is the normalizing constant and Pr (N) is our prior distribution for N. One important possibility for Pr (N) is the improper uniform, which reflects the idea that any particular value of N up to ∞ seems as plausible as any other (subject to the caveat that it must be greater than or equal to the maximum of the sample). Höhle and Held (2006) use binomial series theory to show that a straightforward closed-form expression then exists for the mean of the resulting posterior: $E(N|x) = k - 1/k - 2(m-1)$ for $k > 2$. In our running example, $\hat{N}_{\text{Bayes}} \approx 68$.

The estimators are of varying complexity and have distinct philosophical origins. Related to this, the estimates they produce may differ, especially in small samples, and it is not immediately obvious that one is "superior" to the others. We assume therefore that readers will be interested in seeing the results from each, and will be comforted if these results are similar across estimators (in the sense that the estimates are not simply a function of "cherry picking" a particular approach). These estimators perform optimally when the observed cables are, on a per embassy-year basis, a random sample of the available reference numbers $1, 2, \ldots N$. Goodman (1954, 104) suggests two easily implemented procedures (see online Appendix A) to establish whether this is indeed the case for a given sample. For what follows, we use a Kolmogorov–Smirnov (KS) test in which every set of embassy-year cable reference numbers is compared with a uniform distribution. If the test does not return a statistically significant p-value (i.e., $p > 0.05$) we cannot reject the null of non-uniformity for that sample. Applying this logic, some 641 embassy-year samples were plausibly uniform; this constitutes 63,853 cables in the data altogether.[7]

Ultimately, we suspect that all the embassies in all the years use the same numbering process, and below we will report both sets of results—for the embassy-years that fulfill the uniform sample tests, and separately for all the data together. When uniform assumptions are not met, estimates should be interpreted with more caution.[8] With this in mind, in addition to the Goodman, Bayes, and MLE estimators, we implement a separate regression-based estimator of population size that is less dependent on uniformity (though it does rely more heavily on assumptions about the expected number of cables created per day being relatively fixed over time). In words, for each embassy-year, for any given date in the Manning sample, we regress the maximum serial number observed upon that date on the numeric day of year (e.g., January 1 is 1 and December 31 is 365) on which it was written. More precisely, within each embassy-year, the relevant equation is $M_t = \beta_0 + \beta_1 \cdot t + \varepsilon$, where M_t is the maximum serial observed on day t, t is the numeric calendar day, and ε is the error. For a 365-day calendar year, the regression-based estimate of the total number of cables produced is then $\hat{N} = \hat{M}_{365} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 365$. When the expected rate

[6] Using m as the estimate for the maximum of a discrete uniform, the probability of seeing any particular element of x is $1/m$; if anything larger than m is used, say $m+1$, that probability—and the value of the likelihood—is reduced. But anything smaller than m is impossible, thus m is $\hat{N}_{\text{MLE}}$.

[7] Readers may be concerned as to the power of the tests of uniformity employed. In online Appendix B, we consider this issue and conclude that for most of our embassy-years, for most departures from the uniform, this is not a concern.

[8] In Appendix C, we provide extensive simulation results exploring the precise role between non-uniformity and the nature of the conclusions we can draw about the estimates.

https://doi.org/10.1093/pan/mpv005 Published online by Cambridge University Press

Table 2 Table shows the total number of cables ($\hat{N}$), as well as the corresponding proportion of leaked cables as estimated by our various approaches for the period 2005–2010

Uniform verified						
		Proportion leaked				
Estimator	*(Total)* $\hat{N}$	*Overall*	*Mean*	*25th*	*Median*	*75th*
MLE	1,051,073	0.061	0.141	0.051	0.109	0.188
Goodman	1,123,982	0.057	0.136	0.047	0.102	0.182
Bayes	1,160,520	0.055	0.135	0.046	0.100	0.180
Regression	1,361,770	0.047	0.093	0.021	0.055	0.131
All cables						
		Proportion leaked				
Estimator	*(Total)* $\hat{N}$	*Overall*	*Mean*	*25th*	*Median*	*75th*
MLE	3,277,647	0.050	0.117	0.039	0.089	0.155
Goodman	3,518,654	0.047	0.113	0.036	0.085	0.152
Bayes	3,564,052	0.046	0.112	0.035	0.084	0.150
Regression	3,733,074	0.044	0.086	0.020	0.054	0.122

Notes. Upper panel refers to embassy-years that meet random uniform sample requirement (total sample size = 63,784); lower panel refers to all embassy-years (total sample size = 163,958). The column "overall" is estimated total share of cables leaked, pooling across all embassies; "mean" and "median" denote the average and median proportions leaked across embassy-years; and "25th" and "75th" mark the 25th and 75th percentiles, respectively.

of cables generated per day is fixed over time, the quantity $\hat{\beta}_1$ closely approximates the daily rate of cable generation. Additional details of this model can be found in online Appendix C.

4 Results

The upper panel of Table 2 displays our results for the sample that meets the uniformity assumption. Each row refers to a different estimator, but our immediate observation is that the estimates are very similar in every case, no doubt in part because we have generally large samples to work with. The column labeled "(total) $\hat{N}$" is the sum of the estimates for $\hat{N}$ for every given embassy-year: i.e., the estimate for Bangkok in 2010 plus the estimate Tokyo in 2009 plus the estimate for New Delhi in 2009, and so on (for all 710 satisfying embassy-years in this sample). This is estimated to be between 1,051,073 and 1,361,770 cables. Of these, the 63,784 cables leaked constitute an estimated proportion of roughly 5–6%. That is, for those embassy-years appearing to satisfy the discrete uniformity assumption, we estimate that the WikiLeaks cable leaks represent approximately 5–6% of all cables sent within the period under study.

From studying the rightmost columns of the table, we see that there is a fair degree of variation in terms of leaking: the mean proportion leaked is around 13%, but for some embassy-years it is much larger or smaller. Recalling that there are no "Top Secret" cables in our data, one explanation for this variation is that documents at that classification level are not evenly distributed over embassy-years (literally, some embassies send more high-level material at certain times than others). Consequently, regardless of the $\hat{N}$ estimated for the embassy-year in question, the numerator in the proportion calculation is lower in some cases than others. Figure 1 plots the proportion leaked for all embassy-years in this sample, with some "interesting" cases highlighted.

If we do not restrict to the random uniform samples, and include all embassy-years, we obtain the results in the bottom panel. Again, these are similar across estimators, and the proportion leaked overall is estimated to be around 5% of all cables.

5 Discussion

We estimate that, overall, approximately 5% of all U.S. diplomatic cables in existence between 2005 and 2010 were disclosed by WikiLeaks. Exactly how "severe" this is in terms of absolute size is

https://doi.org/10.1093/pan/mpv005 Published online by Cambridge University Press

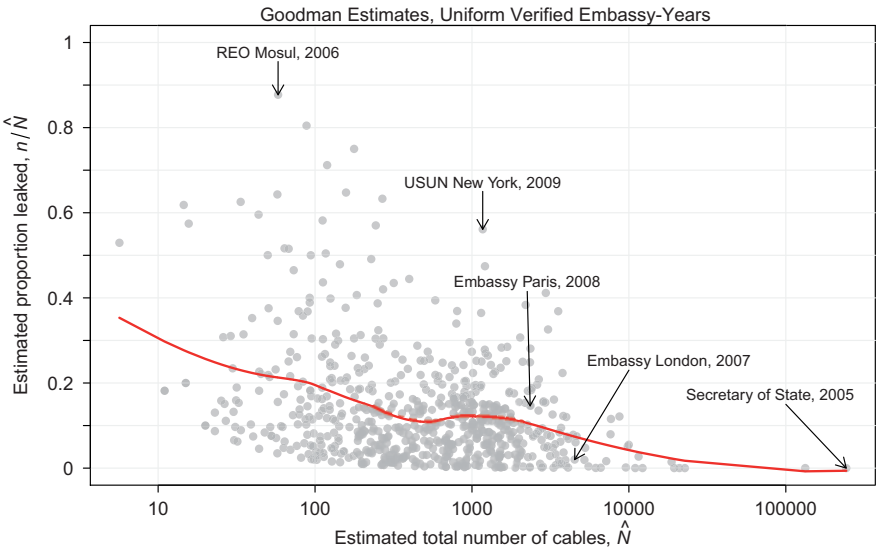


Fig. 1 Plot compares the estimated share of leaked documents at the embassy-year level against the estimated total number of documents in an embassy-year. A solid LOESS regression line is plotted, and the horizontal axis uses a log-10 scale. Overall there is a negative association between the size of the embassy, as measured in total cables produced, and the proportion of cables leaked. Although untestable, a story consistent with this association is that larger embassies have on average higher rates of "sensitive material" produced than smaller embassies (i.e., communications inaccessible to Private Manning, given her security clearance).

debatable. Obviously, the damage done to U.S. interests is at least as dependent on the contents of the cables as it is on the sheer volume of the leak, and it is only the latter that we have investigated here. Generally, our Letter provides methods by which researchers can be aware of the sample size available to them, overall and by embassy-year, should they choose to work with such data. Assuming the official sources (either leaked or declassified) use numbering conventions as discussed here, the same techniques can be used for other projects. Future work might consider the content of the cables, and utilize our estimates in regression-type analyses.

References

Christensen, Thomas. 2006. Fostering stability or creating a monster? The rise of China and U.S. policy toward East Asia. *International Security* 31:81–126.

Durbin, James. 1961. Some methods for constructing exact tests. *Biometrika* 48(1):41–65.

Fearon, James. 1994a. Domestic political audiences and the escalation of international disputes. *American Political Science Review* 88:577–92.

———. 1994b. Signaling versus the balance of power and interests: An empirical test of a crisis bargaining model. *Journal of Conflict Resolution* 38:236–69.

———. 1995. Rationalist explanations for war. *International Organization* 49:379–414.

Gibbs, David. 1995. Secrecy and international relations. *Journal of Peace Research* 32:213–28.

Gill, Michael, and Arthur Spirling 2015. Replication data for: Estimating the severity of the WikiLeaks United States diplomatic cables disclosure. Dataverse [Distributor]. URL: http://dx.doi.org/10.7910/DVN/28729.

Goemens, Hein. 2000. *War and punishment: The causes of war termination and the first world war*. Princeton, NJ: Princeton University Press.

Goodman, Leo. 1954. Some practical techniques in serial number analysis. *Journal of the American Statistical Association* 49:97–112.

Goodman, Leo A. 1952. Serial number analysis. *Journal of the American Statistical Association* 47:622–34.

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58:83–90.

Höhle, Michael, and Leonhard Held 2006. *Bayesian estimation of the size of a population*. Technical Report Discussion Paper 499, Ludwig-Maximilians-Universität München.

Lake, David. 2010. Two cheers for bargaining theory: Assessing rationalist explanations of the Iraq war. *International Security* 35:7–52.

https://doi.org/10.1093/pan/mpv005 Published online by Cambridge University Press

Larsen, Richard, and Morris Marx. 2011. *Introduction to mathematical statistics and its applications*. Upper Saddle River, NJ: Pearson.
Lewis, Peter. 1965. Some results on tests for Poisson processes. *Biometrika* 52:67–77.
Ogata, Yosihiko. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83:9–27.
Ramsay, Kristopher. 2008. Settling it on the field: Battlefield events and war termination. *Journal of Conflict Resolution* 52:850–79.
Reed, William. 2003. Information and economic interdependence. *Journal of Conflict Resolution* 47:54–71.
Reiter, Dan. 2009. *How wars end*. Princeton, NJ: Princeton University Press.
Ruggles, Richard, and Henry Brodie. 1947. An empirical approach to economic intelligence in World War II. *Journal of the American Statistical Association* 42:72–91.
Sagar, Rahul. 2013. *Secrets and leaks: The dilemma of state secrecy*. Princeton, NJ: Princeton University Press.
Schultz, Kenneth. 1999. Do democratic institutions constrain or inform? Contrasting two institutional perspectives on democracy and war. *International Organization* 53:233–66.
Shapiro, Jacob, and David Siegel. 2010. Is this paper dangerous? Balancing secrecy and openness in counterterrorism. *Security Studies* 19:66–98.
Slantchev, Branislav. 2006. Politicians, the media, and domestic audience costs. *International Studies Quarterly* 50:445–77.
Smith, Alastair. 1998. International crises and domestic politics. *American Political Science Review* 92:623–38.
Smith, Alistair, and Allan Stam. 2004. Bargaining and the nature of war. *Journal of Conflict Resolution* 48:783–813.
Snyder, Jack, and Erica D. Borghard. 2011. The cost of empty threats: A penny, not a pound. *American Political Science Review* 105:437–56.
Werner, Suzanne. 1999. The precarious nature of peace: Resolving the issues, enforcing the settlement, and renegotiating the terms. *American Journal of Political Science* 43(3):912–34.

https://doi.org/10.1093/pan/mpv005 Published online by Cambridge University Press