Topic Modeling Official Secrecy

David Allen, Matthew Connelly, Daniel Krasner, Ian Langmore, Thomas Nyberg

Last year, American government officials and private contractors decided to classify information over 80 million times, up from 23 million in 2008.[1] The President's Public Interest Declassification Board estimates that just one intelligence agency produces a petabyte of data yearly, and that it would take two million archivists working full-time to review all of it for declassification.[2] In fact, the government is spending less than half as much money on declassification as it did fifteen years ago.[3] Understaffed archives report having to destroy 95-97% of State Department documents. This includes all of its diplomatic cables related to government-sponsored scientific research not cross-referenced with another subject deemed to have more historical significance.[4]

The growth of official secrecy in the world's only superpower should concern humanists, scientists, and engaged citizens everywhere. Over five million Americans are now required to have security clearances, and the yearly cost of protecting classified information has reached $11 billion -- four billion more than the budget for the National Science Foundation.[5] Such a vast apparatus of secrecy is inimical to any pursuit of knowledge. As historian of science Peter Galison argues, science aims to uncover and secure information, whereas classification makes it impossible for us to know even what we do not know. Secrecy stifles innovation and protects wasteful or misconceived R&D spending. It can also leave research with real value to society unused and unknown.[6] Even when research funding is unclassified, scientists can find their work being used in ways they might never have anticipated. DARPA and IARPA played a key role in catalyzing research in natural language processing and machine-learning, for instance, technologies that are now at the core of both computer science research and cultural concern about surveillance.[7] Even if scientists are uninterested in what U.S. officials deliberate and decide on in secrecy, these officials may well be interested in them. And if even historians are unable to reconstruct their deliberations, American policymakers may never be fully accountable for their actions.

To be sure, some secrets must be safeguarded. This includes the fruits of scientific research that could be massively destructive if they were to fall into the wrong hands. Yet official reviews have repeatedly found that the vast and impossibly complex system created to keep secrets actually makes it harder to identify and protect information that really does pose significant risks.[8] Private watchdog groups have found many examples, such as technical information from chemical and biological weapons programs sitting on the open shelves at the U.S. National Archives.[9]

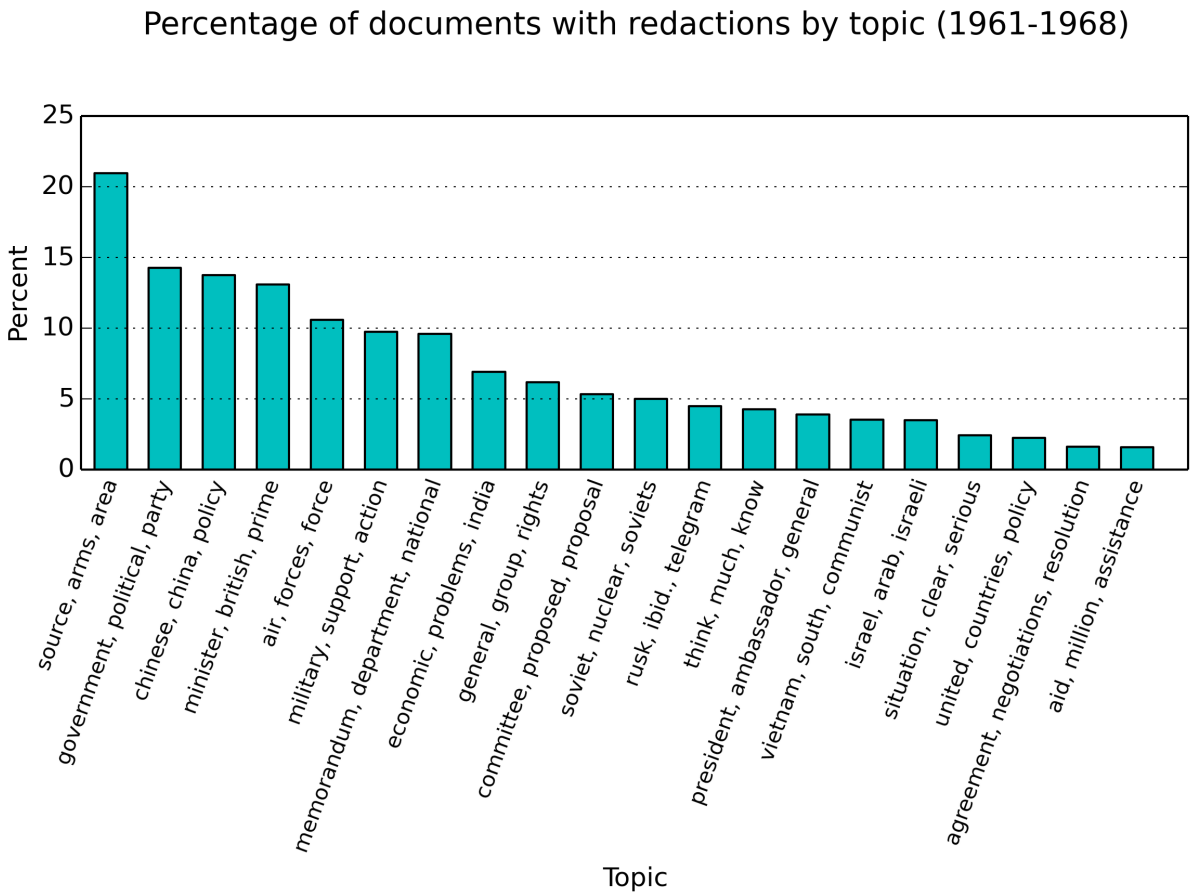
With less and less of the historical record being released to the public, it becomes all the more important to use what tools we have, and to develop new ones, to understand the expanding scope and changing nature of official secrecy. Latent Dirichlet Allocation (LDA) topic modeling is a powerful statistical framework which aims to describe the dynamics of given observations via the interactions of certain hidden or unobserved groups of words, or, more technically, probability distributions over words.[10] In the case of documents, these unobserved groups can be

thought of as "topics" and these topics reflect the semantic distribution of terms. Starting with a corpus of documents, and a way to generate tokens (e.g., words, stems, bigrams), an LDA topic model can be built for the given collection and the words within it. The result is a set of topics/topic probabilities which describe the corpus, such that each document can be allocated a score to represent its association with each topic, and the documents themselves naturally grouped together based on topic affinity. The LDA model thus overlays a rich structure on top of any collection, pointing to different applications as well as avenues for further exploration.

In addition, the topic scores given to every document can work as features in a classification scenario. We have seen this work well, either alone or when combined with metadata or semantic features. For example, we analyzed a collection of 1.1 million declassified State Department cables from the years 1973-1976.[11] We found it was possible to construct a "secrecy" classifier to predict, or score, whether a given State Department cable was originally labeled as secret. Although this project is in its first stages, the preliminary results are strong, with an area under the ROC curve of > 0.9 (where a score of .5 and 1 represent fully random and 100% accurate, respectively) . Further refinements may help identify cables that are misclassified, whether at a higher or lower classification level than the intrinsic sensitivity of the information would appear to warrant.

Using topic modeling, we can also identify the kinds of subjects that are more likely to be redacted, which can be seen as an indicator that they remain sensitive decades after the fact. For instance, we analyzed a curated collection of declassified records from the *Foreign Relations of the United States,* which have been chosen by State Department historians as the most important or representative U.S. foreign policy documents. We split the documents into two periods: 30,184 documents from 1952 to 1960, or the Eisenhower years, and 20,828 documents from 1961 to 1968, corresponding to the administrations of Kennedy and Johnson. We eliminated words that showed up too often or too rarely, including "words," "lines," "paragraphs," "not," "declassified" (the specific words used to indicate redacted text). We then identified twenty topics for each period, and contrasted the most and least redacted topics.

For the 1952-60 period, the group of documents with the largest percentage of redactions is most strongly associated with the words"oil, day, man, times, companies, arabia, construction." Many of the documents most representative of this topic concern the CIA-backed overthrow of the democratically-elected government of Guatemala in 1954. What is more surprising is the prevalence of documents about U.S. oil companies. Some relate to their concern about the emergence of OPEC, but also how "their local man" would help the CIA eliminate Guatemala's oil supplies (often measured in "barrels per day.") Taken together, these documents were 24.3 times more likely to be redacted than the topic with the smallest percentage of redactions, which are most strongly associated with "foreign, exchange, bank, department, export, market, grant." These documents concern the relatively less sensitive work of the Export-Import Bank, and less strategic commodities like Uruguayan wool tops.

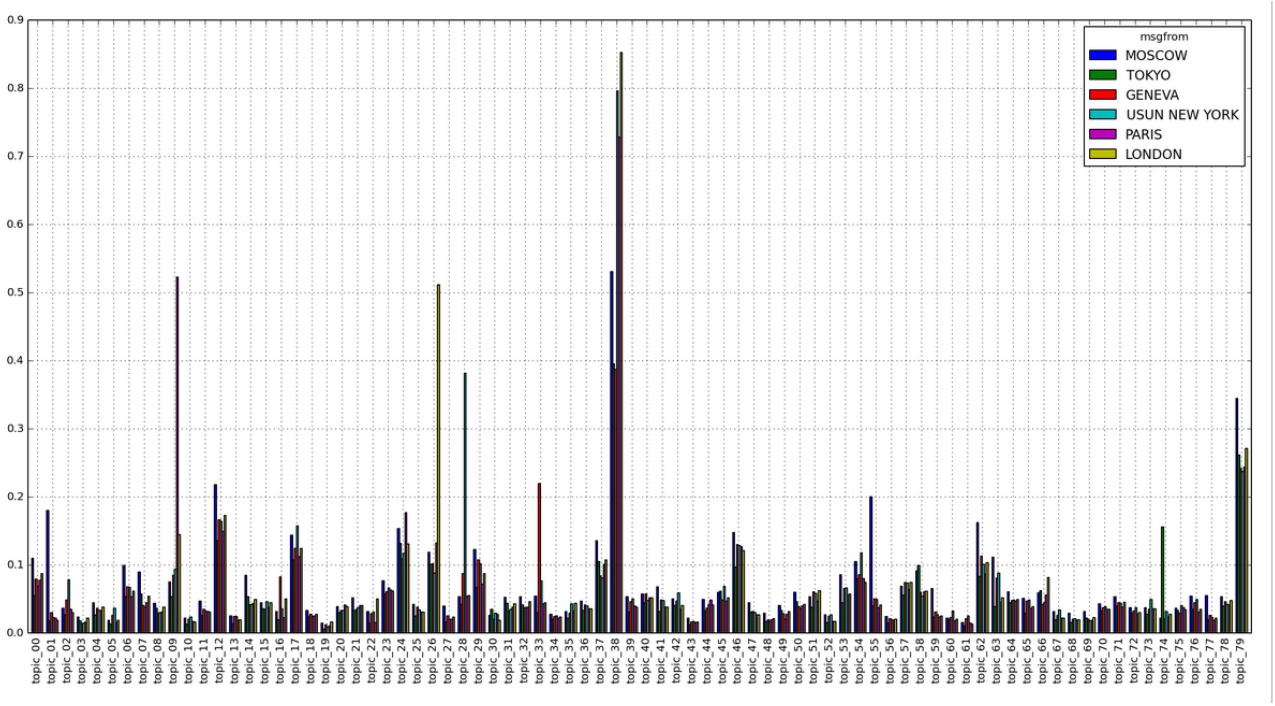


For the Kennedy and Johnson years, the topic with the largest percentage of redactions is most strongly associated with "source, arms, area, mission, information, officers, base." Here again, topic modeling succeeds in grouping documents that are particularly sensitive, such as the covert bombing campaign of Laos, a neutral country in the Vietnam War, and U.S. and Soviet arms shipments to Africa. This topic is 13.3 times more likely to be redacted than documents most strongly associated with "aid, million, assistance, economic, countries, japan, foreign," showing how documents concerning foreign aid and trade are consistently less likely to require redactions. Considering that there are twenty-five times more documents representing this topic than "source, arms, area," this technique also makes it possible to focus on the comparatively small number of documents on topics that officials deem most sensitive.

This research is exploratory in nature, but it allows us to begin to imagine automated techniques whereby officials themselves could prioritize the records they need to scrutinize more closely and accelerate the release of everything else. To be sure, how information is classified and later declassified can be idiosyncratic and context specific. Even a more rational, risk management approach to official secrecy would have to be concerned with disclosures that, however rare, could be unusually damaging. But there are patterns in official communications, and some of them are predictable. This could help us detect anomalies that might otherwise be overlooked.

Communications from each embassy, for instance, have a distinctive signature, with cable traffic

tending to share a specific distribution of words relating to the topics of most concern to that embassy. With over a million diplomatic cables, one can train a text classifier to predict from where a cable originated. Some embassies, like Moscow, were highly predictable in the 1970s, and one can accurately classify these cables 98% of the time. Others, like London, had more visitors passing through and more varied kinds of business, yielding accurate predictions only four times out of five. But these misclassified cables are actually the ones we are looking for, since they reveal when government officials become unpredictable by going "off-topic," engaging in back-channel negotiations, diplomatic gambits, or off-color conversations.



A crucial advantage to these techniques is that they can be scaled up with little additional effort to process even larger corpora -- orders of magnitude larger. To be sure, the input data must be prepared, code written, and parameters specified. But otherwise, minimal human intervention is required to identify both sensitive subjects and anomalous communications. These features will be essential for any system that can cope with millions and eventually billions of e-mails, text messages, etc.

Some secrets need to be kept secure, of that there should be no doubt. Yet when everything is secret, nothing is secret, as recent leaks have made clear. Both citizens and state officials need help in identifying and classifying different kinds of information in large-scale corpora that will otherwise overwhelm traditional declassification methods, and topic modeling may be part of the solution. Ironically, the data-mining techniques that were first developed for intelligence gathering may now provide the only means to keep the government more accountable for the secrets it still keeps.

[1] Information Security Oversight Office (ISOO), "2013 Report to the President" (June 2014), archives.gov/isoo/reports/2013-annual-report.pdf, p.1.
[2] Report to the President from the Public Interest Declassification Board (PIDB), "Transforming the Security Classification System" (November 2012), archives.gov/declassification/pidb/recommendations/transforming-classification.pdf, p. 17.
[3] ISOO, "2013 Report," p. 21.
[4] David Langbart, William Fischer, and Lisa Roberson, "Appraisal of records covered by N1-59-07-3-P" (June 4, 2007).
[5] Brian Fung, "5.1 million Americans have security clearances," *The Washington Post* (March 24, 2014), washingtonpost.com/blogs/the-switch/wp/2014/03/24/5-1-million-americans-have-security-clearances-thats-more-than-the-entire-population-of-norway/; ISOO "2013 Report," p. 20; National Science Foundation, "FY 2015 Budget Request to Congress" (2014), nsf.gov/pubs/2014/nsf14041/nsf14041.pdf.
[6] Peter Galison, "Removing Knowledge," *Critical Inquiry* 31 (2004), pp. 229-243.
[7] Ashton Anderson, Dan McFarland, Dan Jurafsky, "Towards a Computational History of the ACL: 1980-2008," *Proceedings of the ACL*, Special Workshop on Rediscovering 50 Years of Discoveries (2012), pp. 13-21.
[8] "Transforming the Security Classification System"; Office of the Director of National Intelligence, "Intelligence Community Classification Guidance Findings and Recommendations Report" (January 2008), fas.org/sgp/othergov/intel/class.pdf; S. Doc. 105-2, "Report of the Commission on Protecting and Reducing Government Secrecy" (December 31, 1997), gpo.gov/fdsys/pkg/GPO-CDOC-105sdoc2/content-detail.html.
[9] National Security Archive, "Declassification in Reverse" (February 21, 2006), gwu.edu/~nsarchiv/NSAEBB/NSAEBB179/.
[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (January 2003), pp. 993-1022.
[11] National Archives and Records Administration, Access to Archival Databases, Record Group 59, State Department Central Foreign Policy Files (1973-1977), aad.archives.gov/aad/series-list.jsp?cat=WR43.