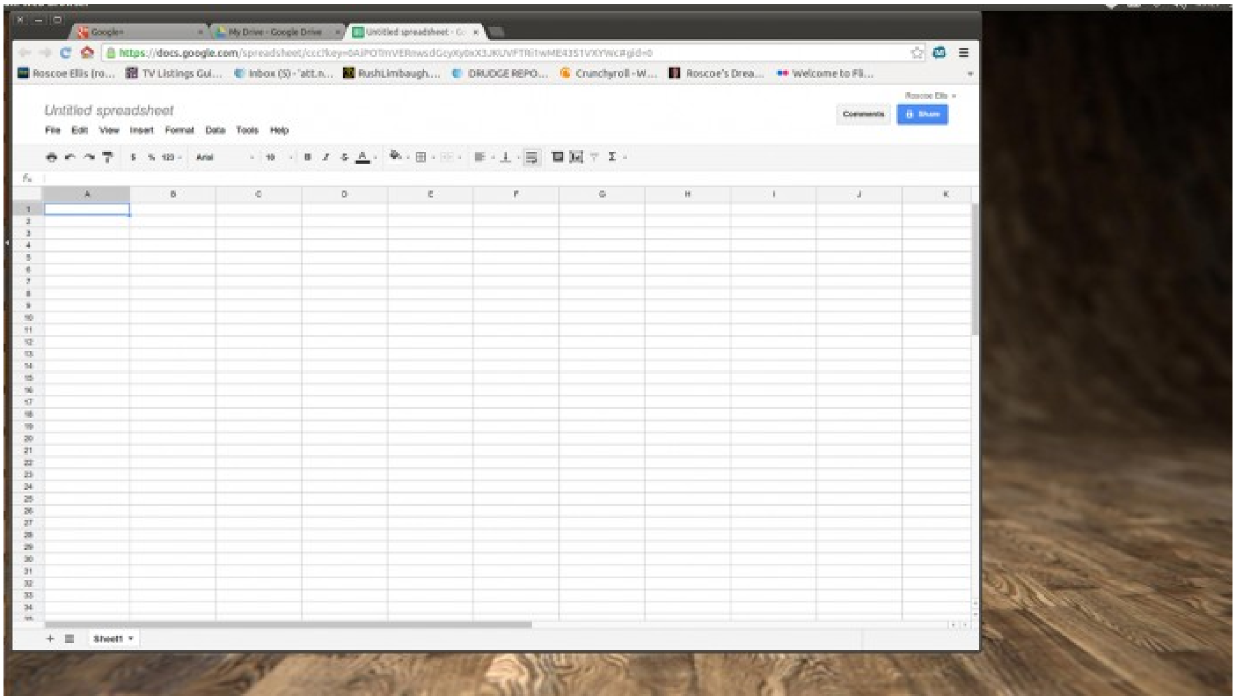
PHYS.ORG

Enron becomes unlikely data source for computer science researchers

April 29 2015, by Matt Shipman



Credit: Roscoe Ellis, shared under a Creative Commons license via Flickr

Computer science researchers have turned to unlikely sources - including Enron - for assembling huge collections of spreadsheets that can be used to study how people use this software. The goal is for the data to facilitate research to make spreadsheets more useful.

"We study spreadsheets because spreadsheet software is used to track everything from corporate earnings to employee benefits, and even simple errors can cost organizations millions of dollars," says Emerson Murphy-Hill, an assistant professor of computer science at NC State and co-author of two new papers on the work.

However, there are relatively few public collections of spreadsheet data available for research purposes. For example, the collection currently used by most researchers consists of approximately 4,500 spreadsheets.

But researchers are now making two new collections available - one has 15,000 spreadsheets and the other has more than 249,000.

"In addition, we are publishing a technique that other researchers can use to collect additional spreadsheet data," Murphy-Hill says.

The 15,000 spreadsheet collection consists entirely of spreadsheets collected from internal Enron emails, which were made public after the emails were subpoenaed by prosecutors.

"Our focus is on how users interact with spreadsheets," Murphy-Hill says. "And these spreadsheets actually tell us a lot about how users represent and manipulate data."

To assemble the second set of spreadsheets, called Fuse, the researchers developed their own technique to identify and extract spreadsheets from an online archive of over 5 billion webpages. Using their technique, the researchers collected 249,376 spreadsheets - including spreadsheets made as recently as 2014.

"Fuse used cloud infrastructure to search through billions of webpages to identify and extract the spreadsheets we write about in this paper," says Titus Barik, a Ph.D. student at NC State, researcher at ABB Corporate Research, and lead author of the paper on Fuse. "Commodity cloud computing is incredibly exciting - searching those pages would take about seven years of continuous computation on a single computer, but the economies of scale with cloud computing allowed us to accomplish this with Fuse in only a few days."

"And the fact that Fuse includes recent spreadsheets is a significant advantage over other spreadsheet collections, because the information is more up-to-date and reflects changes in Excel and other spreadsheet software," Murphy-Hill says.

"Fuse is also more reproducible than other spreadsheet collections," says Kevin Lubick, a Ph.D. student at NC State and co-author of a paper about Fuse. "Reproducibility is the cornerstone of good scientific research, but many existing spreadsheet collections are difficult to reproduce. Our technique can be used by anyone, and they'll get the same results we get. But the results will also include any new spreadsheets made available since the last time the program was run."

More information: The Enron collection is the subject of a paper called "Enron's Spreadsheets and Related Emails: A Dataset and Analysis," which is being presented at the International Conference on Software Engineering May 20-22 in Florence, Italy. Lead author of the paper is Felienne Hermans of Delft University of Technology.

The Fuse paper, "Fuse: A Reproducible, Extendable, Internet-scale Corpus of Spreadsheets," is being presented at the Working Conference on Mining Software Repositories, May 16-17, in Florence, Italy. The Fuse paper was co-authored by NC State Ph.D. students Justin Smith and John Slankas.

Provided by North Carolina State University

Citation: Enron becomes unlikely data source for computer science researchers (2015, April 29) retrieved 24 April 2024 from https://phys.org/news/2015-04-enron-source-science.html