Information loss to public networks

Peter van Bolhuis, Jan-Willem Selij

UNIVERSITEIT VAN AMSTERDAM

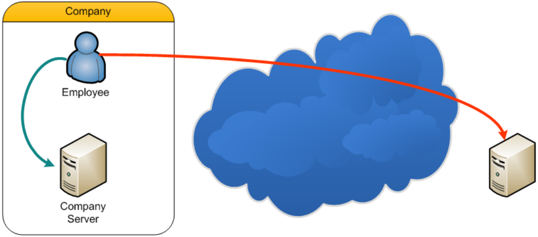
February 4, 2014

ABN·AMRO

1. Intro
2. Research questions
3. Expected results
4. Methods and approach
5. Experiments and data gathering
6. Findings and Analysis
   - Proxy Data
   - Keywords
   - Google Hacking
   - Other Online Detection
7. Conclusion

Intro

- Solutions that allow users to create, modify and share files on the internet have greatly increased the past few years.
- Require little to no integration, easy to use, publicly accessible.
- But also makes it easier to store data on a server one or a company doesn't own...



Research questions

- How can confidential company data efficiently be detected on the most popular services, based on extracted company usage information?
- What online services pose the highest risk for loss of confidential data?

There has been done some related work, namely on the prevention of data loss and watermarking information.

Expected results

Confidential data loss to:

- File sharing sites
  - Such as Dropbox, Mediafire, 4Shared, Zippyshare, ...
- Text sharing sites
  - Such as Pastebin, Paste2, Pastie, dpaste, ...
- Social media
- Office in the cloud
  - Such as Office 365, Google Docs, Evernote, Prezi...
- Personal storage, such as a NAS

Methods and approach

- Determine most-used services
- Determine keywords and identifiers for data
- Develop a method for searching each used service type

Experiments and data gathering

- ABN AMRO offered us some insight to the proxy logs for about 23,500 employees.
    - Of which 80% are Dutch.
    - Representable for large Dutch companies.
    - Created a query to run daily, for a week

    - Only requests >50KiB
    - Created a script to aggregate this data

Proxy Data

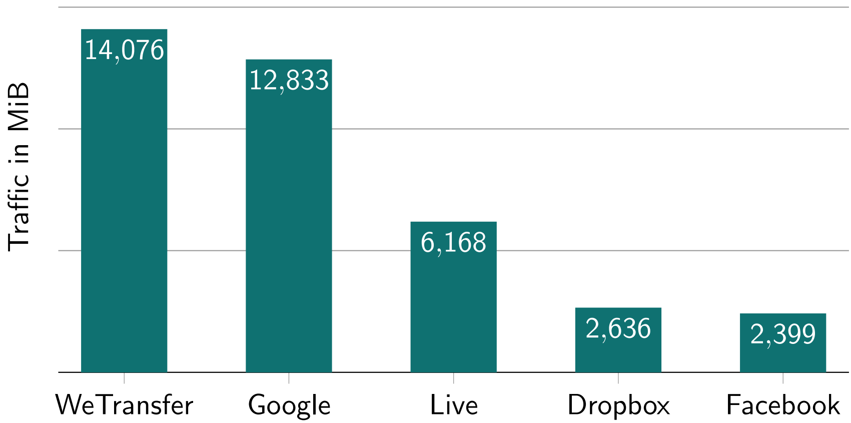


Figure : Outgoing network traffic aggregated by domain (1 week)

Proxy Data

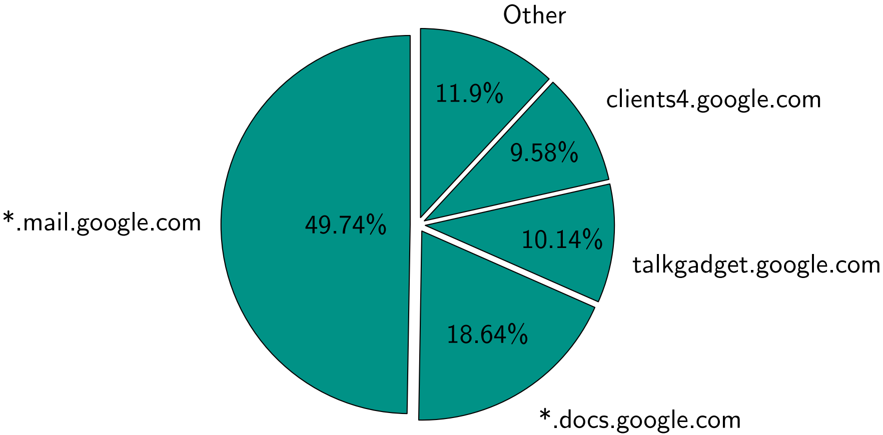


Figure : Google sub-domains

Keywords

- Using data from DLP system

- Looking at disclaimers of confidential files

Keywords

- Using data from DLP system
  - Proprietary
  - Trade secret
  - Internal use only
  - Not for distribution
  - Various terms specific to departments
  - Various regular expressions
- Looking at disclaimers of confidential files

Keywords

- Using data from DLP system
    - Proprietary
    - Trade secret
    - Internal use only
    - Not for distribution
    - Various terms specific to departments
    - Various regular expressions
- Looking at disclaimers of confidential files
    - Confidential
    - Classified
    - Strictly Personal

Google Hacking

- Logical operators

- Special operands

Google Hacking

- Logical operators
  - NOT
  - AND
  - OR
  - Grouping ()
- Special operands

Google Hacking

- Logical operators
  - NOT
  - AND
  - OR
  - Grouping ()
- Special operands
  - filetype:
  - inurl:
  - intext:
  - ...

Google Hacking Results

- filetype:doc | filetype:docx | filetype:pdf AND ("abn amro" OR "abnamro") AND (-inurl:abn OR -inurl:abnamro) "overgemaakt op rekeningnummer *" "Sofinummer"
- filetype:doc | filetype:txt | filetype:pdf AND ("abn amro" OR "abnamro") AND (-inurl:abn OR -inurl:abnamro) vertrouwelijk

Google Hacking Results

- filetype:doc | filetype:docx | filetype:pdf AND ("abn amro" OR "abnamro") AND (-inurl:abn OR -inurl:abnamro) "overgemaakt op rekeningnummer *" "Sofinummer"
  - **Pension and Salary information**
- filetype:doc | filetype:txt | filetype:pdf AND ("abn amro" OR "abnamro") AND (-inurl:abn OR -inurl:abnamro) vertrouwelijk

Google Hacking Results

- filetype:doc | filetype:docx | filetype:pdf AND ("abn amro" OR "abnamro") AND (-inurl:abn OR -inurl:abnamro) "overgemaakt op rekeningnummer *" "Sofinummer"
    - **Pension and Salary information**
- filetype:doc | filetype:txt | filetype:pdf AND ("abn amro" OR "abnamro") AND (-inurl:abn OR -inurl:abnamro) vertrouwelijk
    - **Confidential documents**

Other Online Detection

- Social networks
- Dropbox
- Cloud Offices
- Other file sharing sites
- Online text-sharing

Conclusion

How can confidential company data efficiently be detected on the most popular services, based on extracted company usage information?

What online services pose the highest risk for loss of confidential data?

Conclusion

How can confidential company data efficiently be detected on the most popular services, based on extracted company usage information?

- Encrypted communication
- Authentication
- Other data can most efficiently be detected with Google

What online services pose the highest risk for loss of confidential data?

Conclusion

How can confidential company data efficiently be detected on the most popular services, based on extracted company usage information?

- Encrypted communication
- Authentication
- Other data can most efficiently be detected with Google

What online services pose the highest risk for loss of confidential data?

- WeTransfer attachments
- Private e-mail