MISP Project

Data Science applied on MISP Data

Mallik, A. Gillard, S.

Data Science in Techno-Socio-Economic Systems
Computer Social Science
ETH Zurich

Project Presentation, May 2022

Table of Contents

1. Introduction
2. Overview
3. Data
4. Problem Statement and Research Questions
5. Method
6. Results
7. Conclusion

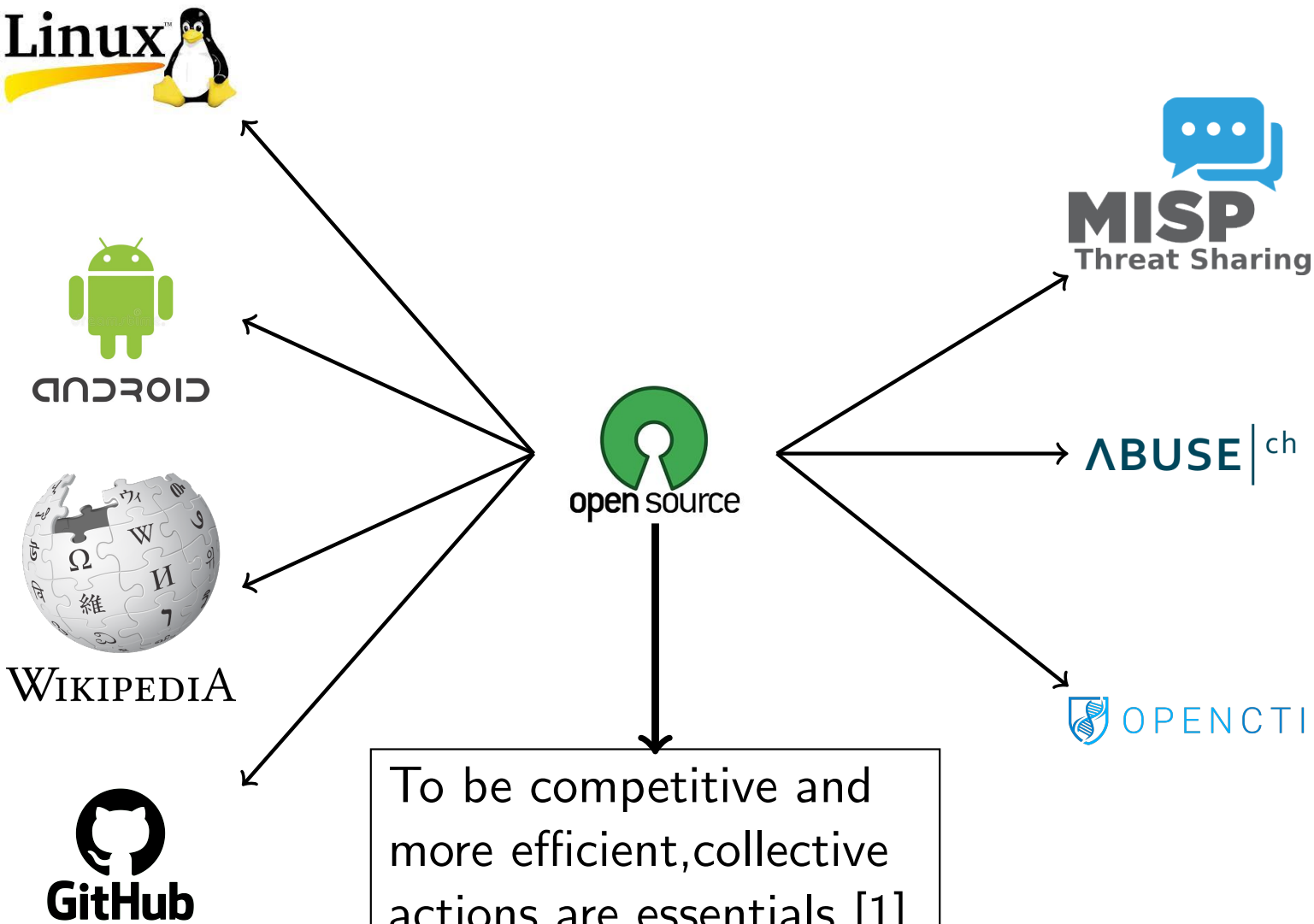
Introduction

Context

- Research field: Cybersecurity - Information-Sharing
- Data set from an information-sharing platform
- Research question:
  - How can we accelerate the response to cyber-threats by harnessing collective intelligence?
  - How shared information is reused down the road to characterize future cyber-threats?

Introduction

Motivation



Overview in the Research Area

Cybersecurity Information-Sharing

- Cyber-criminals have sharing experiences $\Rightarrow$ success
- Information asymmetry between attackers and defenders [2]
- Threat intelligence platform, as an open-source solution:
  - aggregation, correlation and analyze threats
  - access to multiple source in real time defensive actions

Overview in the Research Area

Collective Action

- Development of collaborative platforms
  - Reduction of the risks of security breaches
  - Highlight the importance of information-sharing
  - Promotion of knowledge transfer
  - Induction of cascades of collective production
- Open collaboration $\Rightarrow$ Production follows a super-linear law
- Key aspect: Knowledge Integration

Overview in the Research Area

Knowledge Integration [3]

- Each individual in a subsystem:
  - brings added value in her own field $\rightarrow$ Differentiation
  - production of a complex good by pooling together these added values $\rightarrow$ Integration
- Knowledge seen as strategic resource $\Rightarrow$ Consideration of specialized knowledge for the organizations to provide competitive advantage
- Creation of more reused knowledge by virtual teams than individual production

Data

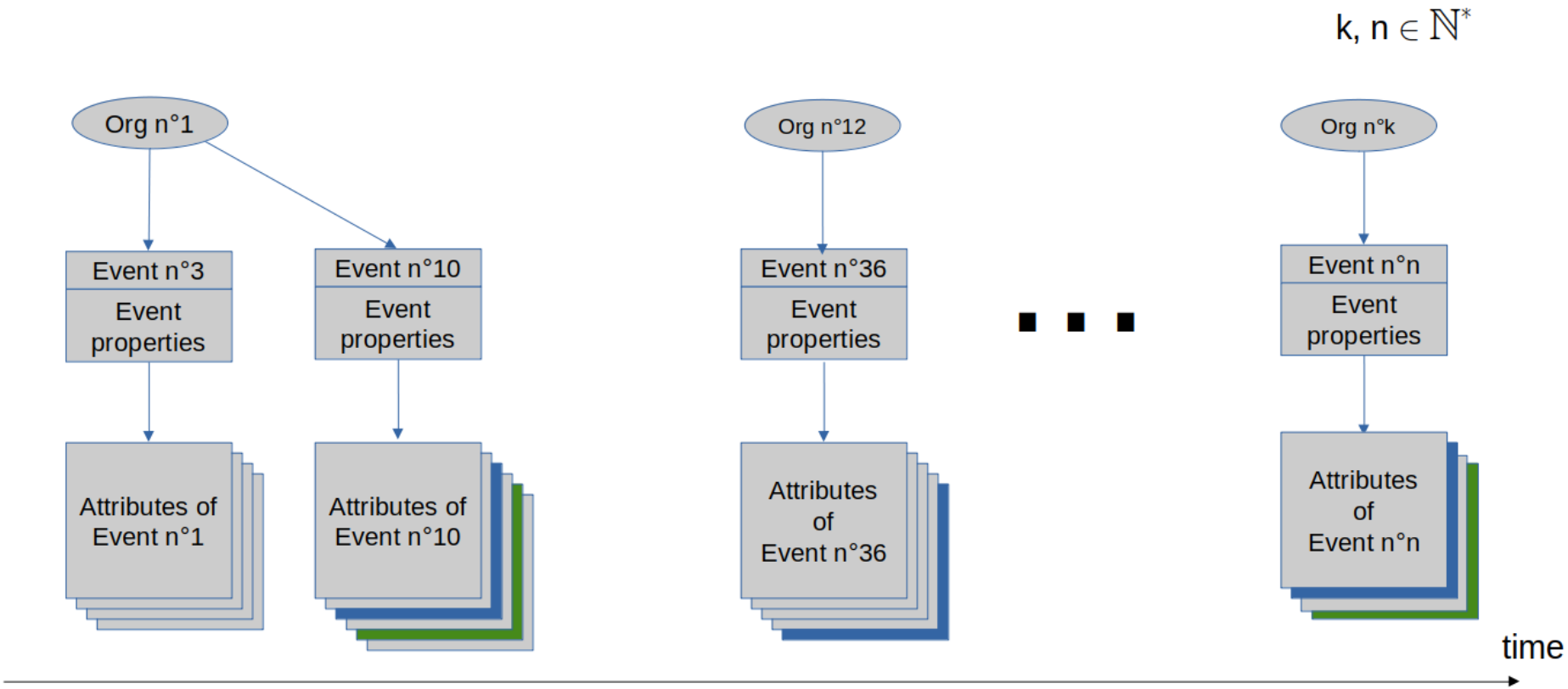
MISP: Malware Information and Threat Sharing Platform [4]

- Popular open-source platform, subdivided in communities
- Create by the Computer Incident Response Center Luxembourg (CIRCL)
- Used by NATO, some governments and organizations
- Offer the possibility to share, stores and collaborate on incidents
- Threats (i.e. events) are characterized by indicators of compromise (i.e. attributes)

Advantages: As an open-source platform, able to study how the platform is designed and works.

MISP: Malware Information and Threat Sharing Platform

Functioning



MISP: Malware Information and Threat Sharing Platform

Raw Data

Event ID	Org ID	Attribute ID	UUID	Attribute Type	Attribute Category	Distribution	Timestamp	Event Date	Event Pub	Publish	Deleted	Value	Comment
78777	1598	12417603	44eeb1a5-eaaf-4d9c-91c8-1129a569dde0	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	chamboncaytrongmarigoldcatbacomwpincludesus_uscorporationinvoice_number344947	emotet url
78777	1598	12417604	bf1bf497-b929-4322-94b4-f011efbcdce1	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	chamboncaytrongmarigoldcatbacomwpincludesus_uscorporationinvoice_number344947	emotet url
78777	1598	12417605	be8db0b4-1bef-4e9c-b2d4-8146f8ae3651	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	cateringbangkokinthwpcontentusscaninvoice_numberkuzfus4_trevkinp	emotet url
78777	1598	12417606	38dc9c7c-b2f4-4aaf-a944-cf3e8396fe60	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	wingmedcomtrdownloadinvoice1334904212119tsapl6u_beyk	emotet url
78777	1598	12417607	4396d8f6-51f2-4e0e-af9c-29f8ee389210	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	youthinenergyorginfoxlqz7b_mvgbte	emotet url
78777	1598	12417608	7ae86a20-5d0f-4e98-a8e7-7abc802d7969	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	youthinenergyorginfoxlqz7b_mvgbte	emotet url
78777	1598	12417609	dc3794c7-4b8f-404b-9e11-ecaf78ce620d	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	wingmedcomtrdownloadinvoice1334904212119tsapl6u_beyk	emotet url
78777	1598	12417610	65139e10-eee9-42f7-be80-0b44c00dc9e7	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	10421122628enfilesgis4tumz_ctftxeuxtr	emotet url
78777	1598	12417611	92a986d5-ee3b-4345-9af4-a6f446614131	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	10421122628enfilesgis4tumz_ctftxeuxtr	emotet url
78777	1598	12417612	c0381e5c-1d23-490f-9f84-89039a51b664	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	cateringbangkokinthwpcontentusscaninvoice_numberkuzfus4_trevkinp	emotet url
78777	1598	12417613	be92d084-64a7-4ab6-8d0d-32de8ab06fc8	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	sciagemeuzacoiscomglqkaymq085sopa	emotet url
78777	1598	12417614	e37922bc-a32c-47b1-808f-f783e23f4093	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	sciagemeuzacoiscomglqkaymq085sopa	emotet url
78777	1598	12417615	77b19c5e-1c65-4a4d-b15e-c1f852412c4d	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	galeriakolashgaleriacollagecomveb8kfy2zfzq4q	emotet url
78777	1598	12417616	028e1574-8bb7-47e0-98f0-76225aab88b3	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	galeriakolashgaleriacollagecomveb8kfy2zfzq4q	emotet url
78777	1598	12417617	3e2a653d-4327-4a79-9b79-1ccf8f2199a8	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	smehelpdesknet80nawj6zjxyj_vjzhhoqas	emotet url
78777	1598	12417618	17f9c34e-2c7c-455b-8104-69b51d6f3cb1	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	smehelpdesknet80nawj6zjxyj_vjzhhoqas	emotet url
78777	1598	12417619	416f3ee8-3d3d-4bd0-af4b-ba932c824a44	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	samaradekorrugbzrcgbbsdngmylc	emotet url
78777	1598	12417620	cc922953-a72c-4c9e-8d3f-cfdcaf231341	url	Network activity	5	1550158171	14/02/2019	1.6E+09	TRUE	FALSE	samaradekorrugbzrcgbbsdngmylc	emotet url

Figure: Data obtained after collection, manipulation and curation

Dataset

- Data collected from November 10th, 2008 to February 8th 2022
- The dataset from the MISP CIRCL instance is constituted of:

Property	Quantity
# of organizations	1,908
# of users	4,013
# of events	39,639
# of attributes	9,099,685
# of tags	3,786
Size of the dataset	14.6 Gb

Table: Properties of the dataset

Data: Collection, Manipulation and Curation

- Automated collection via PyMISP and the Rest API $\rightarrow$ `.json`
  - 39,639 `.json`, i.e. 1 per event
- Selection of the parameters and transformation into `.csv`
- Curation:
  - Transformation and standardization of data (int, float, str, ...)
  - Remove the org n° 1203: Abusiv dump of this org

Problem Statement and Research Questions

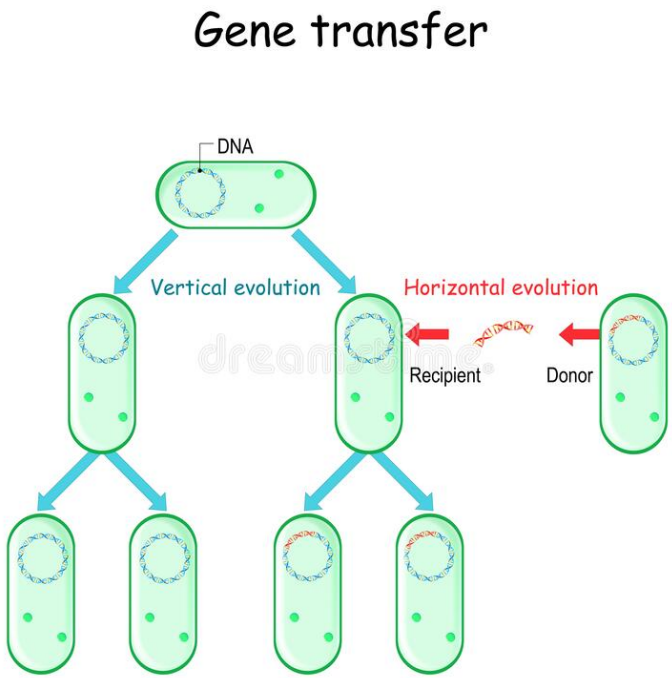


Figure: Vertical and Horizontal Gene Transfer [5].

- How shared information is reused down the road to characterize future cyber-threats?

Method

Summary

- Dichotomic Evaluation of Inheritance
  - Importance and Links between the Events
  - Distribution of Mothers and Daughters
- NLP & Vectorization
- Complex Network of Attributes

- **Comparison of the "Value" of the attributes**
  - Attribution of an ID per "Value"
  - Condition:

$$\text{Value ID} = \begin{cases} \text{if str are } 100\% \text{ similar, Value ID are similar,} \\ \text{if str are different, Value ID are different.} \end{cases}$$

  - The same "Value ID" give information about the link between the events
    - Undirected Graph
  - The links between the events are carried with the attributes

Method

Dichotomic Evaluation of Inheritance

- Complementary Cumulative Distribution Function of Mothers per Daughter
- Fit Method: Transform in log-log and make a linear regression $\Rightarrow$ Power Law [6]

Method

NLP & Vectorization

- **Normalization** with NLP:
  - URL: if value begins with 'https://www.', truncate it
  - Remove punctuation in all attribute types except IP-addresses
  - Convert all values to lower case

Method

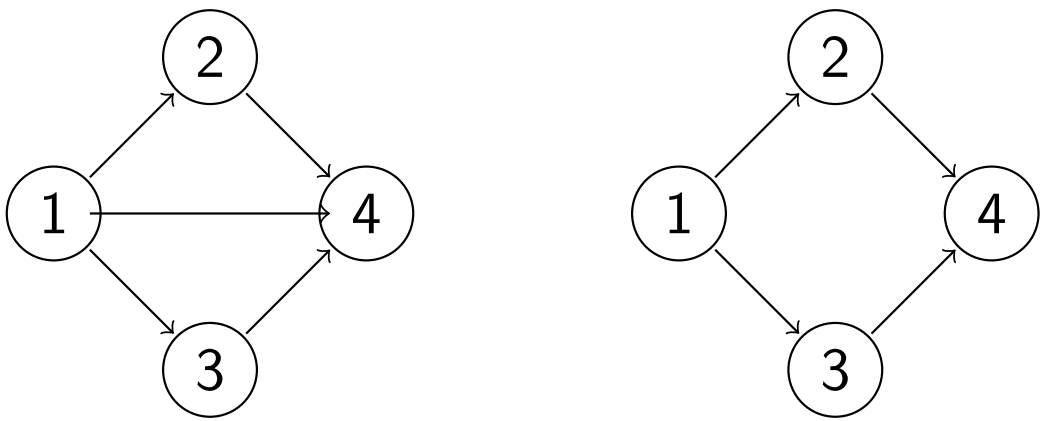
NLP & Vectorization

- **Normalization** with NLP:
  - URL: if value begins with 'https://www.', truncate it
  - Remove punctuation in all attribute types except IP-addresses
  - Convert all values to lower case
- **Vectorization**:
  - *Inheritance*: similarity ratio between 2 values
  - `SequenceMatcher.ratio()` from Python's `difflib` library
  - I matrix with inheritance between 2 attribute values as entries $\rightarrow$ $(I)_{i,j} = 2M/T$ is inheritance of Att_j's value from Att_i's [7]
  - I is symmetric but in context of chronology it is a lower triangular matrix

Method

Weighted Directed Graph between Attributes

- **Transformation** of the vectorization matrix into a graph
  - Row and column indexes give the edges
  - Elements give the weight of the edges
  - To know the direction of the edges, comparison of the creation timestamps for the corresponding attributes.
- **Transitive reduction** (see scheme below)
- Use of `Python` library: `networkx`



Results

Importance and Links between the Events

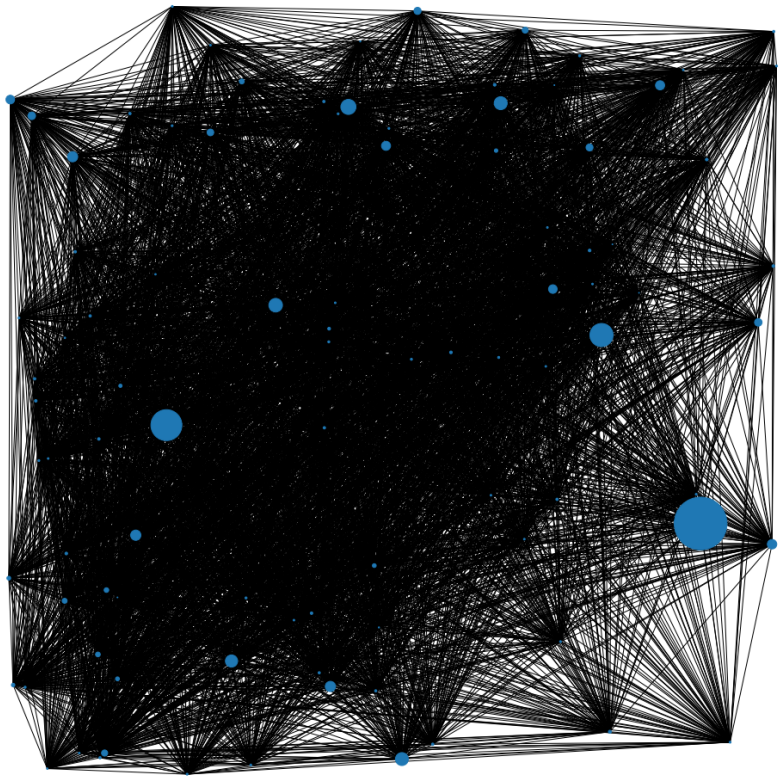


Figure: Undirected graph representing the edges between the events. The nodes size represents the number of attributes encapsulated in the corresponding event. To do this figure, 10000 rows were selected.

Results

Importance and Links between the Events

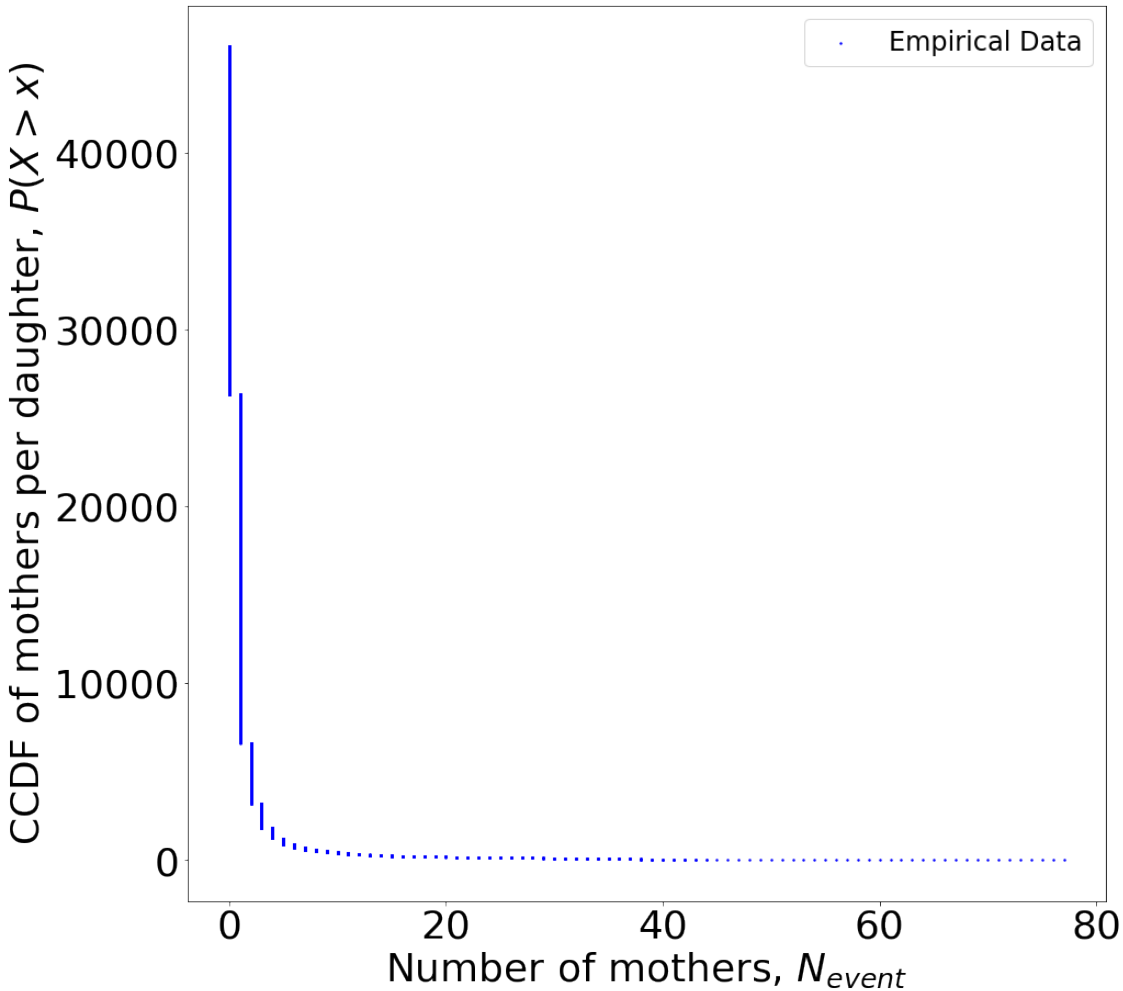


Figure: Complementary cumulative distribution function (CCDF) of mothers per daughter. The data follow an heavy-tailed distribution.

Event ID	Attribute ID	Value
44371	6468741	http://fastchem.co.id/muri/config.bin
44371	6468742	http://fastchem.co.id/muri/bot.exe
44371	6468743	http://fastchem.co.id/muri/gate.php
44371	6468744	http://fastchem.co.id/kays/config.bin
44371	6468745	http://fastchem.co.id/kays/bot.exe
44371	6468746	http://fastchem.co.id/kays/gate.php
44371	6468747	103.28.15.136
44371	6468748	fastchem.co.id

Table: Raw values of Event ID 44371

Results

Normalized Values

Event ID	Att_i	Attribute ID	Value
44371	1	6468741	fastchemcoidmuriconfigbin
44371	2	6468742	fastchemcoidmuribotexe
44371	3	6468743	fastchemcoidmurigatephp
44371	4	6468744	fastchemcoidkaysconfigbin
44371	5	6468745	fastchemcoidkaysbotexe
44371	6	6468746	fastchemcoidkaysgatephp
44371	7	6468747	103.28.15.136
44371	8	6468748	fastchemcoid

Table: Normalized values of Event ID 44371. Normalization done with NLP and Python's `re` library.

Results

Vectorization Example

	Att_1	Att_2	Att_3	Att_4	Att_5	Att_6	Att_7	Att_8
Att_1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Att_2	0.7	1.0	0.0	0.0	0.0	0.0	0.0	0.0
Att_3	0.7	0.8	1.0	0.0	0.0	0.0	0.0	0.0
Att_4	0.8	0.5	0.6	1.0	0.0	0.0	0.0	0.0
Att_5	0.6	0.8	0.7	0.7	1.0	0.0	0.0	0.0
Att_6	0.5	0.6	0.8	0.7	0.8	1.0	0.0	0.0
Att_7	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Att_8	0.6	0.7	0.7	0.6	0.7	0.7	0.0	1.0

Results

Directed Graph

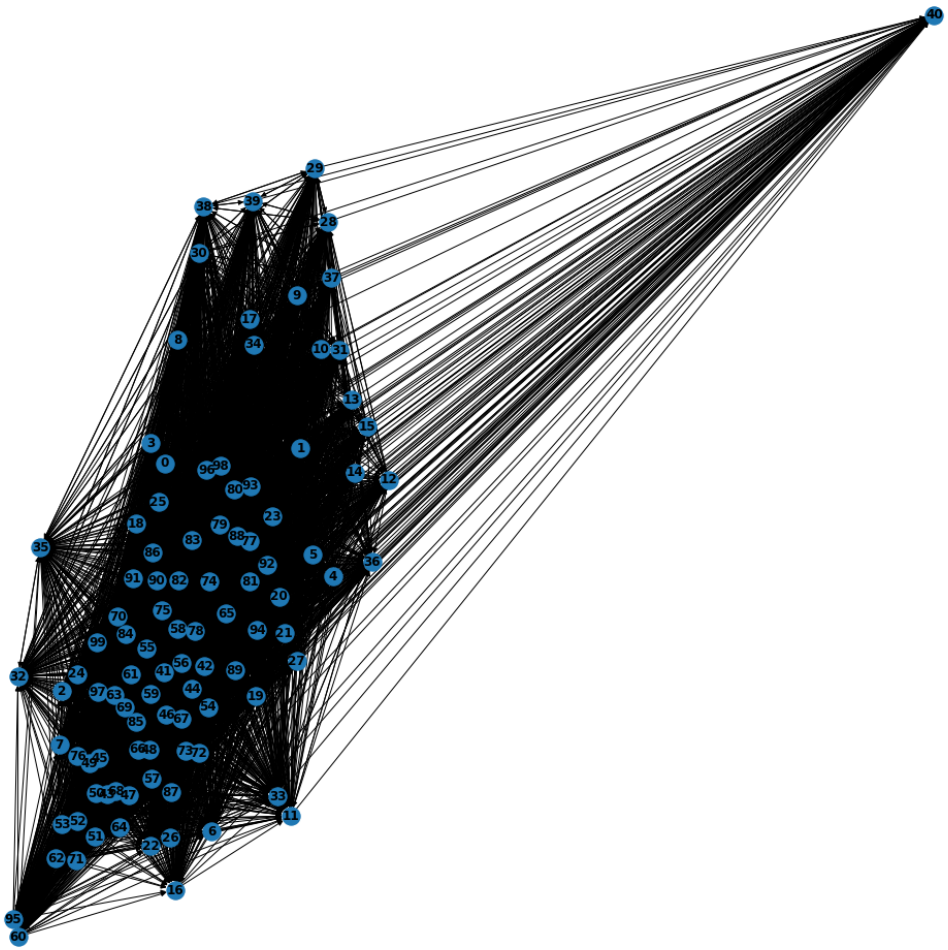


Figure: Directed graph of the links between attributes based on the vectorization matrix. A subset of 100 nodes is used here.

Conclusion

Thus, we were able to address the research question: there is reuse of information and it occurs via horizontal transfer.

Issues

- Mechanisms of information reuse (i.e. inheritance between attributes) were more complex than expected
- Resulted in run-time for code often being much longer than expected

Importance

- Humans are the center of information-sharing
- Platforms like MISP allow us to transmit relevant information
- This allows decisions to be made efficiently to resolve problems quickly
- A good understanding of such information-sharing mechanisms will result in better algorithms

Questions

References I

[1] *Open Source Software and the "Private-Collective" Innovation Model: Issues for Organization Science*, en. DOI: 10.1287/orsc.14.2.209.14992. [Online]. Available: https://pubsonline.informs.org/doi/epdf/10.1287/orsc.14.2.209.14992 (visited on 05/16/2022).

[2] S. Laube and R. Böhme, "Strategic Aspects of Cyber Risk Information Sharing," *ACM Computing Surveys*, vol. 50, no. 5, 77:1–77:36, Nov. 2017, ISSN: 0360-0300. DOI: 10.1145/3124398. [Online]. Available: https://doi.org/10.1145/3124398 (visited on 05/12/2022).

References II

[3] D. Engel and T. W. Malone, "Integrated information as a metric for group interaction," en, *PLOS ONE*, vol. 13, no. 10, C. Dovrolis, Ed., e0205335, Oct. 2018, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0205335. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0205335 (visited on 05/16/2022).

[4] C. Wagner, A. Dulaunoy, G. Wagener, and A. Iklody, "MISP: The Design and Implementation of a Collaborative Threat Intelligence Sharing Platform," en, in *Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security*, Vienna Austria: ACM, Oct. 2016, pp. 49–56, ISBN: 978-1-4503-4565-1. DOI: 10.1145/2994539.2994542. [Online]. Available: https://dl.acm.org/doi/10.1145/2994539.2994542 (visited on 03/29/2022).

References III

[5] M. Juhas, J. R. van der Meer, M. Gaillard, R. M. Harding, D. W. Hood, and D. W. Crook, "Genomic islands: Tools of bacterial horizontal gene transfer and evolution," en, *FEMS Microbiology Reviews*, vol. 33, no. 2, pp. 376–393, Mar. 2009, ISSN: 1574-6976. DOI: 10.1111/j.1574-6976.2008.00136.x. [Online]. Available: https://academic.oup.com/femsre/article-lookup/doi/10.1111/j.1574-6976.2008.00136.x (visited on 05/12/2022).

[6] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," en, *SIAM Review*, vol. 51, no. 4, pp. 661–703, Nov. 2009, arXiv:0706.1062 [cond-mat, physics:physics, stat], ISSN: 0036-1445, 1095-7200. DOI: 10.1137/070710111. [Online]. Available: http://arxiv.org/abs/0706.1062 (visited on 05/16/2022).

References IV

[7] *Difflib — Helpers for computing deltas — Python 3.10.4 documentation*, [Online]. Available: https://docs.python.org/3/library/difflib.html (visited on 05/15/2022).

References of images I

1. Linux: http://www.lespace.co.jp/en/topics/2016/02/19/fastcopy-for-linux-open-source-project-started/
2. Android: https://trendblog.net/the-story-of-the-android-logo/
3. Wikipedia: https://fr.m.wikipedia.org/wiki/Fichier:Wikipedia-logo-v2-wordmark.svg
4. GitHub: https://logos-world.net/github-logo/
5. Open Source: https://icon-icons.com/icon/opensource-logo/169884
6. MISP: https://www.misp-project.org/
7. Abude.ch: https://abuse.ch/
8. OpenCTI: https://www.opencti.io/fr/

References of images II

9. Gene transfer: https://fr.dreamstime.com/bact%C3%A9ries-transfert-g%C3%A8nes-exemple-%C3%A9volution-horizontale-verticale-donneur-cellule-%C3%A0-destinataire-contact-image196348166
10. Human: https://www.google.com/search?q=human+logo+black+and+white&tbm=isch&ved=2ahUKEwiH1OjUy-P3AhUMGRoKHbqWD2wQ2-cCegQIABAA&oq=human+logo&gs_lcp=CgNpbWcQARgGMgQIIxAnMgUIABCABDIECAAQHjIECAAQHjIECAAQHjIECA
sclient=img&ei=1gmCYoeREoyyaLqtvuAG&bih=906&biw=1920&client=firefox-b-lm#imgrc=il93LHKOjGYCmM
11. Algorithm: https://www.dreamstime.com/algorithm-icon-vector-sign-symbol-isolated-white-backgrou

References of images III

12. Decision: https://fr.dreamstime.com/ic%C3%B4ne-vecteur-prise-d%C3%A9cision-d-affaires-image137929220
13. Questions: https://fernelmont.ecolo.be/2018/05/27/conseil-communal-18-avril-2018-projets-brouillons-a-ferne always-ask-questions-logo/