Text Mining Workshop GROUP 10

Customers Complaints about Financial Services and Products

Bellomo Michele, De Mario Giorgia, Galatà Serena, Monte Federica, Rendine Francesca, Schino Pasquale, Sgobba Matteo, Summo Emanuela

FUNDAMENTAL ASPECTS

DATASET USED

The *Consumer Complaint Database* is a collection of complaints about consumer financial products and services, published by the **CFPB (Consumer Financial Protection Bureau)** in 2022.

Complaints are published after the company has responded, confirming a business relationship with the consumer, or after 15 days, whichever comes first.

cfpb Consumer Financial Protection Bureau

OBJECTIVE OF THE ANALYSIS

Identify, through a text mining analysis and the subsequent visualization of the results, the **main problems** and **critical issues** reported in the complaints for some companies representative of the initial "population".

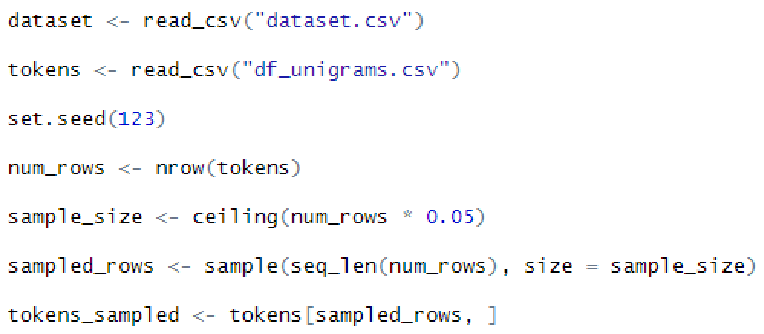
The text mining techniques we use include: **frequency analysis, tf-idf, coword and cosine similarity.**

How can text mining transform customer complaints into opportunities for business improvement?

STEP 1: DATA PREPARATION AND CLEANING

Dataset reduction

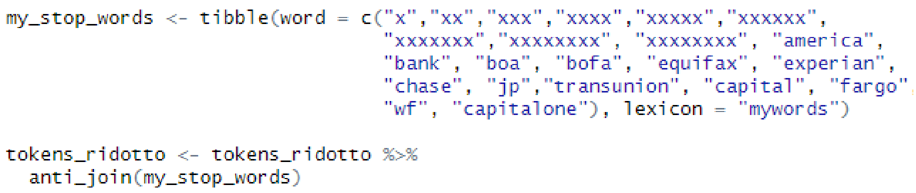
Given the considerable size of the token dataset, we decided to use a sample reduced to **5%** of the total data:



```
dataset <- read_csv("dataset.csv")

tokens <- read_csv("df_unigrams.csv")

set.seed(123)

num_rows <- nrow(tokens)

sample_size <- ceiling(num_rows * 0.05)

sampled_rows <- sample(seq_len(num_rows), size = sample_size)

tokens_sampled <- tokens[sampled_rows, ]
```

Remove stopwords, numbers and unnecessary words

We then removed **stopwords**, **numbers** and all those **words that were not useful** for the analysis:

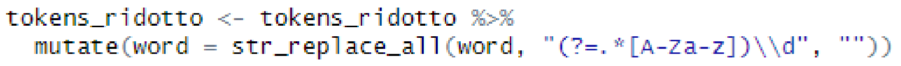


```
my_stop_words <- tibble(word = c("x","xx","xxx","xxxx","xxxxx","xxxxxx",
                                 "xxxxxxx","xxxxxxxx", "xxxxxxxx", "america",
                                 "bank", "boa", "bofa", "equifax", "experian",
                                 "chase", "jp","transunion", "capital", "fargo",
                                 "wf", "capitalone"), lexicon = "mywords")

tokens_ridotto <- tokens_ridotto %>%
  anti_join(my_stop_words)
```

Lemmatisation and use of RegEx

To facilitate the analysis and make it more meaningful, we applied lemmatisation, reducing the words to their base form.

Furthermore, with a **Regular Expression** we were able to remove numbers from all words that contained numbers:



```
tokens_ridotto <- tokens_ridotto %>%
  mutate(word = str_replace_all(word, "(?=.*[A-Za-z])\\d", ""))
```

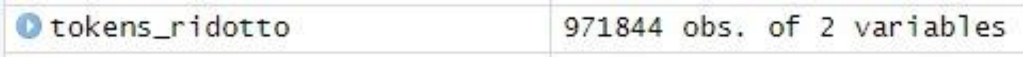
OBSERVATION

Using only 5% of the total data, most of the obtained plots (mainly frequencies and TF-IDF) demonstrated the same results that would have been obtained by using the full dataset, obviously with frequencies scaled accordingly.

STEP 2: SAMPLING

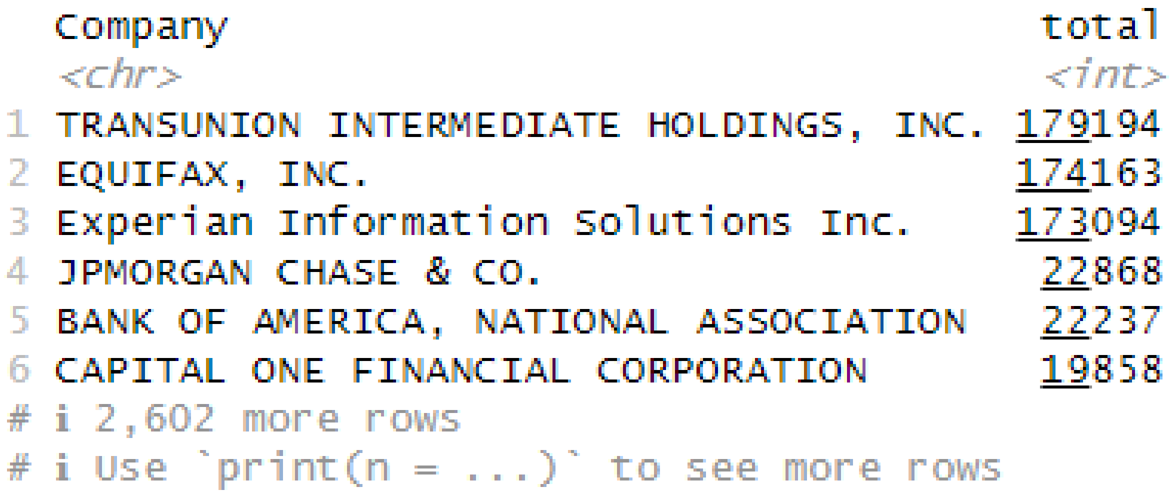
The six companies with the highest number of tokens in the dataset were selected to provide a representative sample for analysis. The companies identified include **TransUnion, Equifax, Experian, JPMorgan Chase & Co., Bank of America, and Wells Fargo**.

These companies represent approximately **61%** of the total dataset.



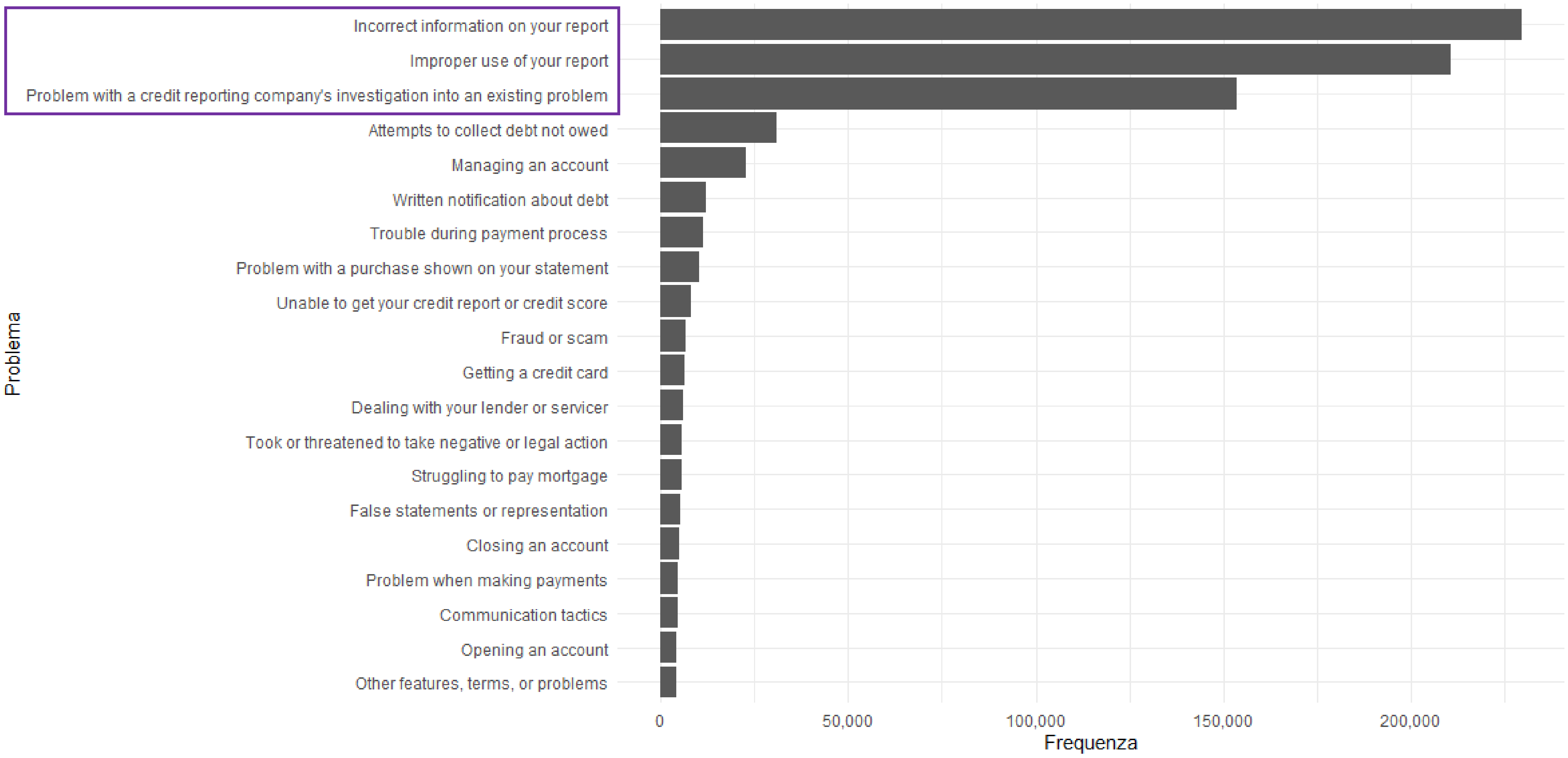
```
tokens_ridotto          971844 obs. of 2 variables
```

Sampling allows us to focus the analysis on the entities with the greatest impact in terms of complaint volume, ensuring greater representativeness of the main trends and critical issues present in the dataset.



```
  Company                                    total
  <chr>                                      <int>
1 TRANSUNION INTERMEDIATE HOLDINGS, INC.    179194
2 EQUIFAX, INC.                             174163
3 Experian Information Solutions Inc.       173094
4 JPMORGAN CHASE & CO.                       22868
5 BANK OF AMERICA, NATIONAL ASSOCIATION      22237
6 CAPITAL ONE FINANCIAL CORPORATION          19858
# ℹ 2,602 more rows
# ℹ Use `print(n = ...)` to see more rows
```

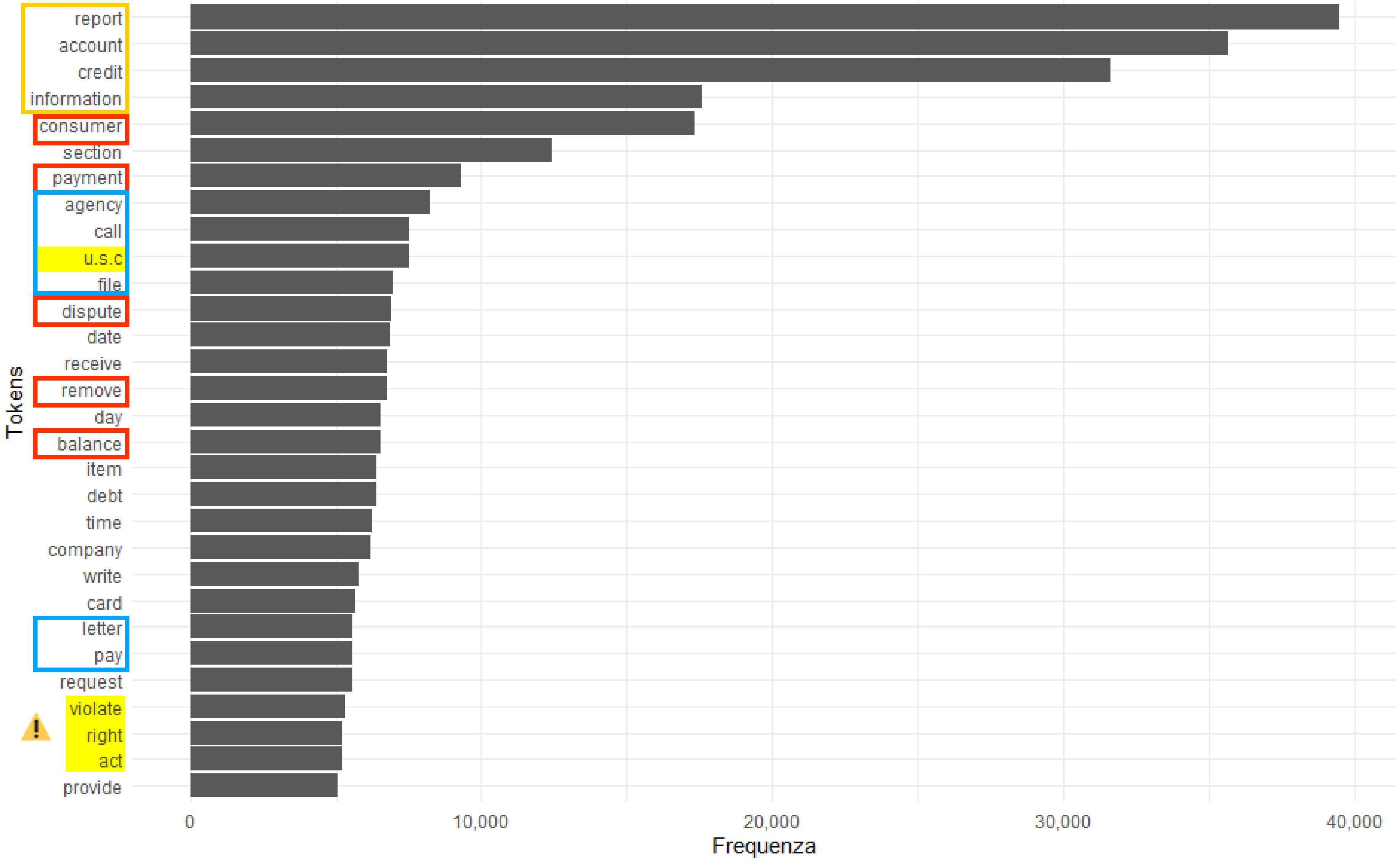
SUPPORT ANALYSIS: UNDERSTANDING THE CONTEXT



The chart shows the main problems encountered by users following some operations requested from companies.
Through the analysis of the frequencies detected, it is possible to highlight how the communication of incorrect information in the reports and the improper use of the latter are the prevalent critical issues.

STEP 3: GLOBAL FREQUENCY ANALYSIS

The following chart shows the most frequent words representing **2%** of the starting sample:



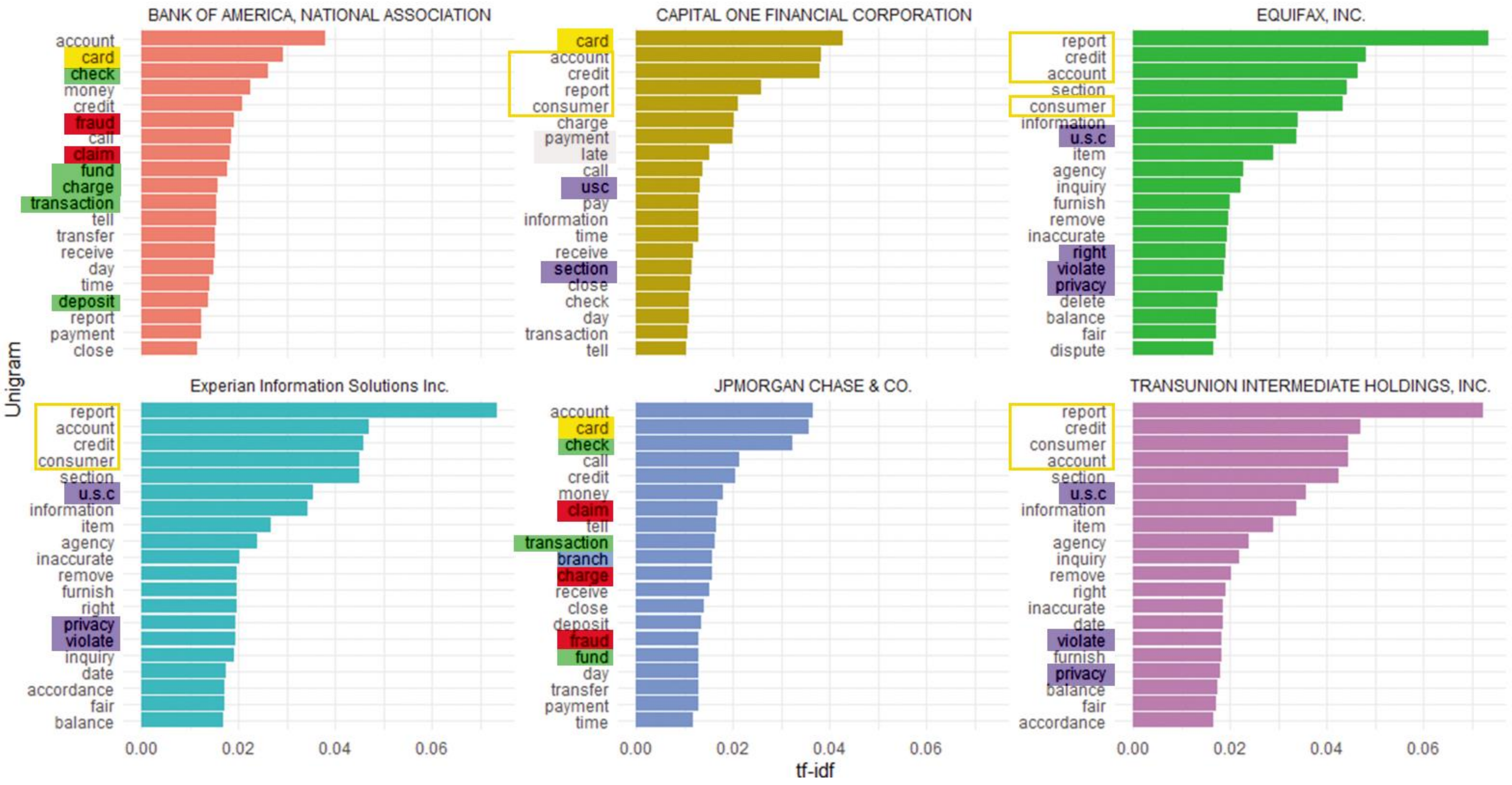
The most recurring words, such as ***'report', 'account', 'credit', 'information'*** suggest a predominance of themes related to the management of financial reports, accounts and credit.

The high frequency of terms such as ***'consumer' , 'payment', 'dispute', 'remove' and 'balance'*** indicate consumer concerns about transactions and disputes to remove incorrect information from reports.

Terms such as ***'agency', 'file', 'call' and 'section'*** suggest a strong involvement of agencies or institutions that manage consumer data, references to specific legal sections (**'u.s.c.'**) and direct communications (**'call'** and **'letter'**).

STEP 4: TF-IDF ANALYSIS

The TF-IDF analysis allowed us to identify more distinctive keywords for the 6 reference companies, revealing the differences between them and excluding irrelevant or generic words:



Legend for highlights:

- ○ Credit Reporting and Balance Sheet Problems
- ● Credit card problems
- ● Fraud, Refunds and Disputed Charges
- ● Traditional banking issues (deposits, transactions and checks)
- ● Regulatory and privacy issues
- ● Delays in payments
- ● Reference to physical branches

Traditional banks (Bank Of America and JPMorgan) mainly present operational and service problems typical of retail banks.
Complaints from **credit agencies** (Equifax, Experian and TransUnion) instead, focus on credit reports, references to regulations and violations of privacy.
Finally, **Capital One** presents intermediate characteristics, with an emphasis on credit cards and payments.

STEP 5: CO-WORD ANALYSIS

Next, we performed a co-word analysis, which allowed us to highlight **correlations between words** within the complaints and therefore **deeper connections** between the problems.

For ease of visualization, and to highlight the most central relationships, it was decided to select **the 30 most frequent words** from the overall dataset of tokens relating to the 6 companies.

The following slide will show **two co-word graphs** compared:

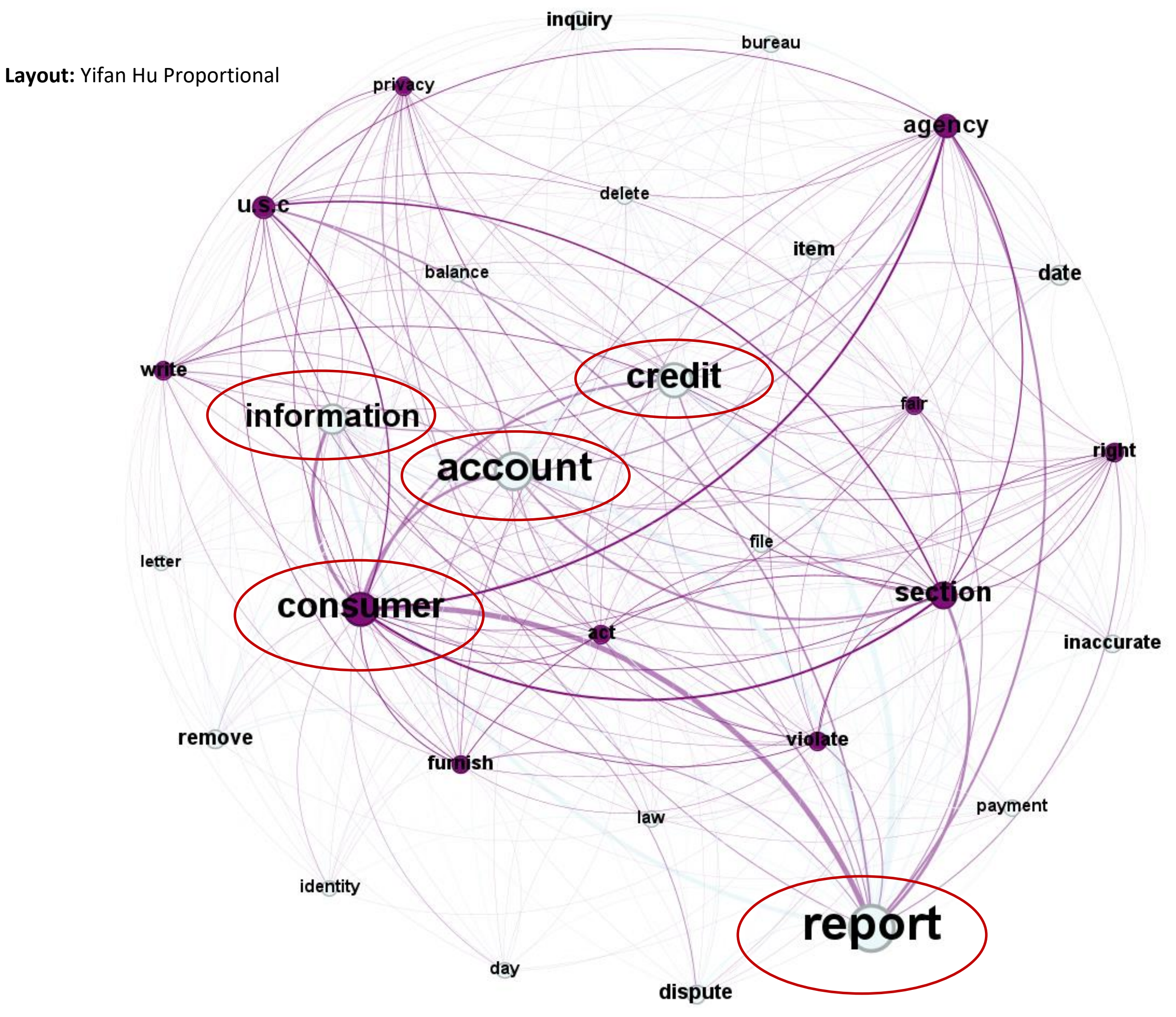
The **first network** represents the 30 most frequent words with their co-occurrence relationships; however, the central nodes are quite generic (although still significant and supporting the previous analyses), and this leads to hiding further more specific correlations.

For this reason, we chose to remove the 5 main words ('credit', 'information', 'account', 'report', 'consumer'), re-plotting a **second network** and highlighting more precisely the connections between more specific and relevant terms.

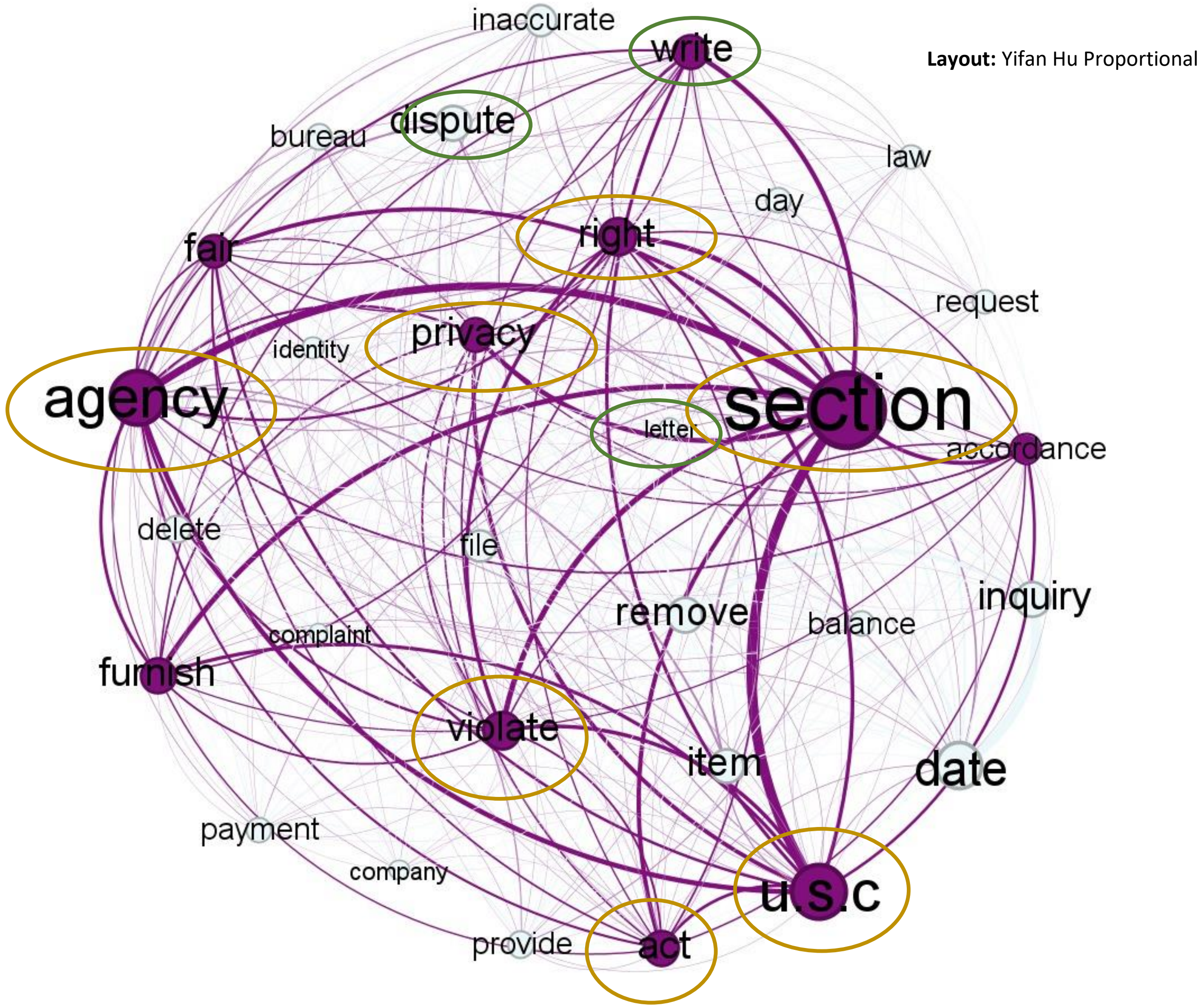
The choice to display the words common to the 6 companies in a single co-word graph guarantees a **competitive advantage** as it allows to evaluate the individual critical issues also in relation to those of your competitors, in such a way as to fully understand the strengths and weaknesses and develop targeted strategies.

STEP 5: CO-WORD ANALYSIS

These networks were generated using a visualization software (Gephi):



Layout: Yifan Hu Proportional

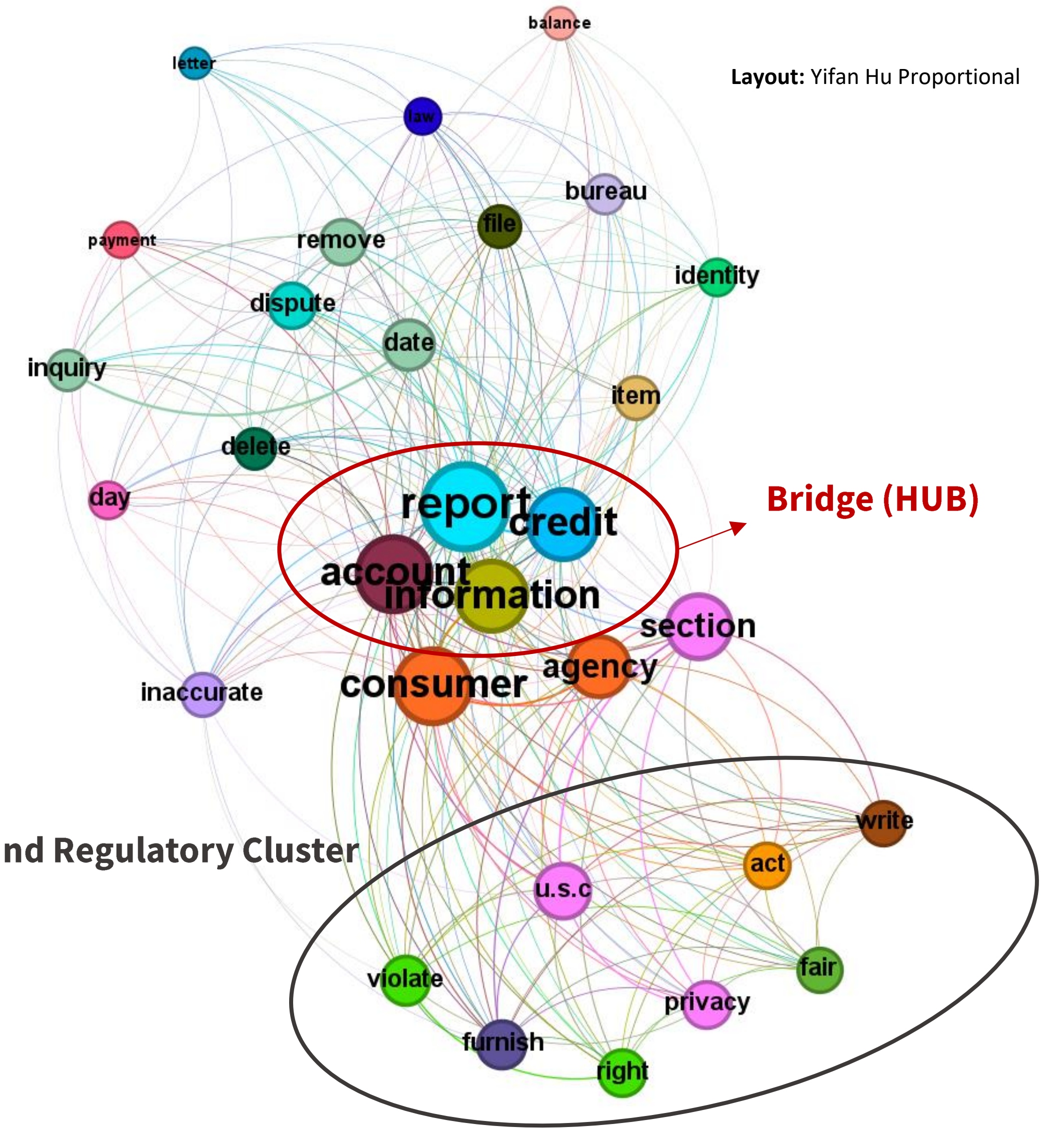


Layout: Yifan Hu Proportional

⚠ Violations of regulations and privacy, with references to sections and acts of the US code and regulatory agencies.

STEP 6: COSINE SIMILARITY

Subsequently, with a cosine similarity analysis, it was possible to evaluate the similarity between words taking into account the length of the complaints and normalizing the data, obtaining a greater number of clusters:



Layout: Yifan Hu Proportional

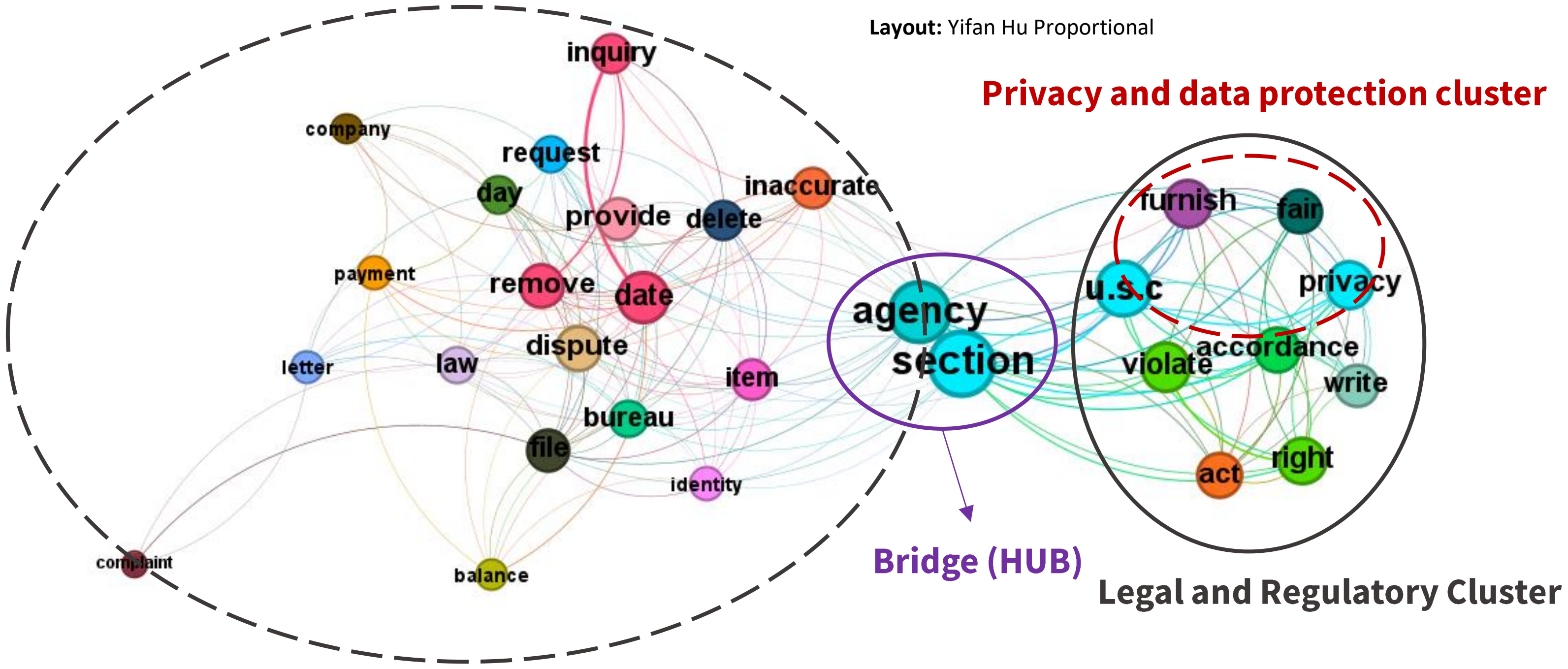
Bridge (HUB)

Legal and Regulatory Cluster

Cosine similarity graph including the 30 most frequent words of the 6 chosen companies.

As with the co-word, token **report**, **credit**, **account**, **information** and **consumer** were found to be the most relevant but quite generic.

We therefore decided to remove them and repeat the analysis, to highlight further correlations:



Layout: Yifan Hu Proportional

Privacy and data protection cluster

Bridge (HUB)

Legal and Regulatory Cluster

Action cluster: remove, delete, date, provide...

CONCLUSIONS AND POSSIBLE INTERVENTIONS

Co-word and **cosine similarity analyses**, as well as **TF-IDF** and **frequency analyses**, provided us with important insights into the complaint trends for the companies we selected.

In light of these observations, a series of possible solutions can be deduced to support corporate decision-making processes:

- Companies should first focus their resources **on improving customer data management**, introducing clearer, safer and more efficient information systems; it would be appropriate to strengthen data controls and optimize the processes of correction and deletion on reports.
- Ensure **proper regulatory compliance**, for example by investing in internal audits and also by modifying policies regarding the processing of personal data and privacy.
- Companies should improve the **security** of their systems, preventing possible cyber attacks and identity theft (and therefore, data).
- **Improve customer service** by providing faster responses to customers who require specific actions.
- In the case of **traditional banks**, it would be necessary to strengthen **anti-fraud processes** and improve the infrastructure of services offered to customers, such as crediting procedures, cheques and money transfers.

Thanks for your attention!

Bellomo Michele

De Mario Giorgia

Galatà Serena

Monte Federica

Rendine Francesca

Schino Pasquale

Sgobba Matteo

Summo Emanuela