OTTAWA 2023

64TH WORLD STATISTICS CONGRESS

DEUTSCHE BUNDESBANK
EUROSYSTEM

IPS 239 - Financial innovation and official statistics

Recognizing Fintechs: A Graph Approach. Automatically Classifying Companies For Statistical Purposes

Andy Bosyi, Neusinger.ai
Ulf von Kalckreuth, Deutsche Bundesbank
Maximilian König, Deutsche Bundesbank

Tuesday 18 July, 2 p.m. - 3:40 p.m.

Fintech and Statistics

- Fintech is the place where **innovation takes place in the financial sector**, where **new methods and products emerge, are tested and made ready for the market.**

- **As yet, fintech is not a separately defined field of statistical activity**, neither in the Bundesbank nor in most other central banks.

- **Three important messages:**

  - Monitoring fintech is **important**
  - Monitoring fintech is **difficult**
  - Monitoring fintech is **feasible**

Monitoring fintech is important!

We need to **monitor fintech**:

- Much of innovation activity in the financial sector takes place in fintechs. Fintechs are **essential for the growth dynamics** of the financial sector.
- Fintechs are of increasing importance for **financial stability**, **supervision**, **payment** and **monetary policy** – the key business areas of central banks.
- **IFC Working Group on Fintech Data Issues**, see final report of July 2020: https://www.bis.org/ifc/publ/ifc_report_monitoring_financial_innovation.pdf

Monitoring fintech is important!

In the short run:

- **Data Gaps Initiative**: New DGI agreed – tough recommendations on fintech
  - Rec 10: Fintech credit
  - Rec 11: Digital Money and Crypto Assets
  - Rec 12: Fintech enabled financial inclusion

- **Classification of activities and products**, ongoing revisions
  - ISIC / NACE
  - CPC / CPA

- **CMFB** has fintech as focus topic

Monitoring fintech is difficult!

- **Few standardised reporting requirements**
- No developed taxonomy and no established set of quantitative measures.
- By definition, innovation involves **new activities**. Intrinsically difficult for traditional statistics, which **needs stable classifications.**
- **Company registers mostly useless**.
- Business environment and market structures are **changing rapidly.**
- High rate of “metabolism” in the sector: **entry, merger & acquisition, exit**
- Any list of fintech companies is rapidly outdated.

We have to do without all the cornerstones of traditional company level statistics

Monitoring fintech is feasible!

- New ways of collecting data need to be explored. **Ways that use mostly publicly available information plus information that is shared voluntarily.**
- Market activity of fintechs are **almost exclusively web-based**. We can **make use of this!**
- Fintech monitoring **needs to be transnational –** firms often do not reside in the country where they are most active.
- To cope with innovation and market dynamics, we **need to become faster!**

We present an graph-based method to detect new fintechs fast

Important for classification tasks in other innovative business areas!

Current stance of project

- Currently, **no follow-up envisaged!**
- Neither for f**intech data collection** nor for **automatised classification.**

What is the use of the POC anyway?

- Strategic planning is carried out in a rolling manner
- Fintech identification and classification world-wide issue, not only in DE, but all other countries
  - Comparable work at BdF, joint paper
  - Approach useful for more general classification purposes

Scope: Objectives and Delimitations

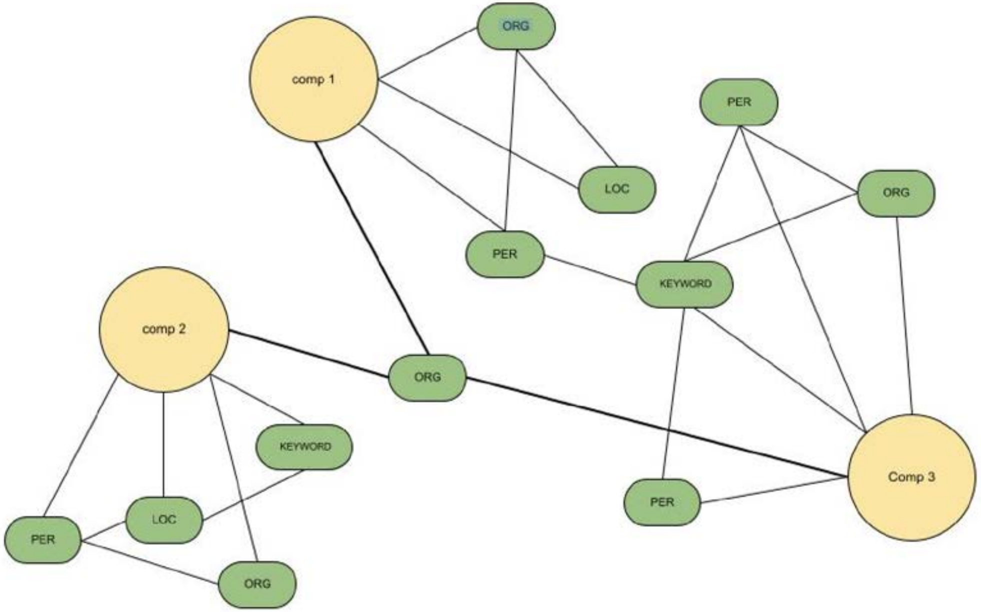
Focus

- Recognising fintechs among IT firms (binary classification)
- Data acquisition via web scraping
- Graph structure as a model of fintech eco system
- Feasibility study

Boundaries / Delimitations

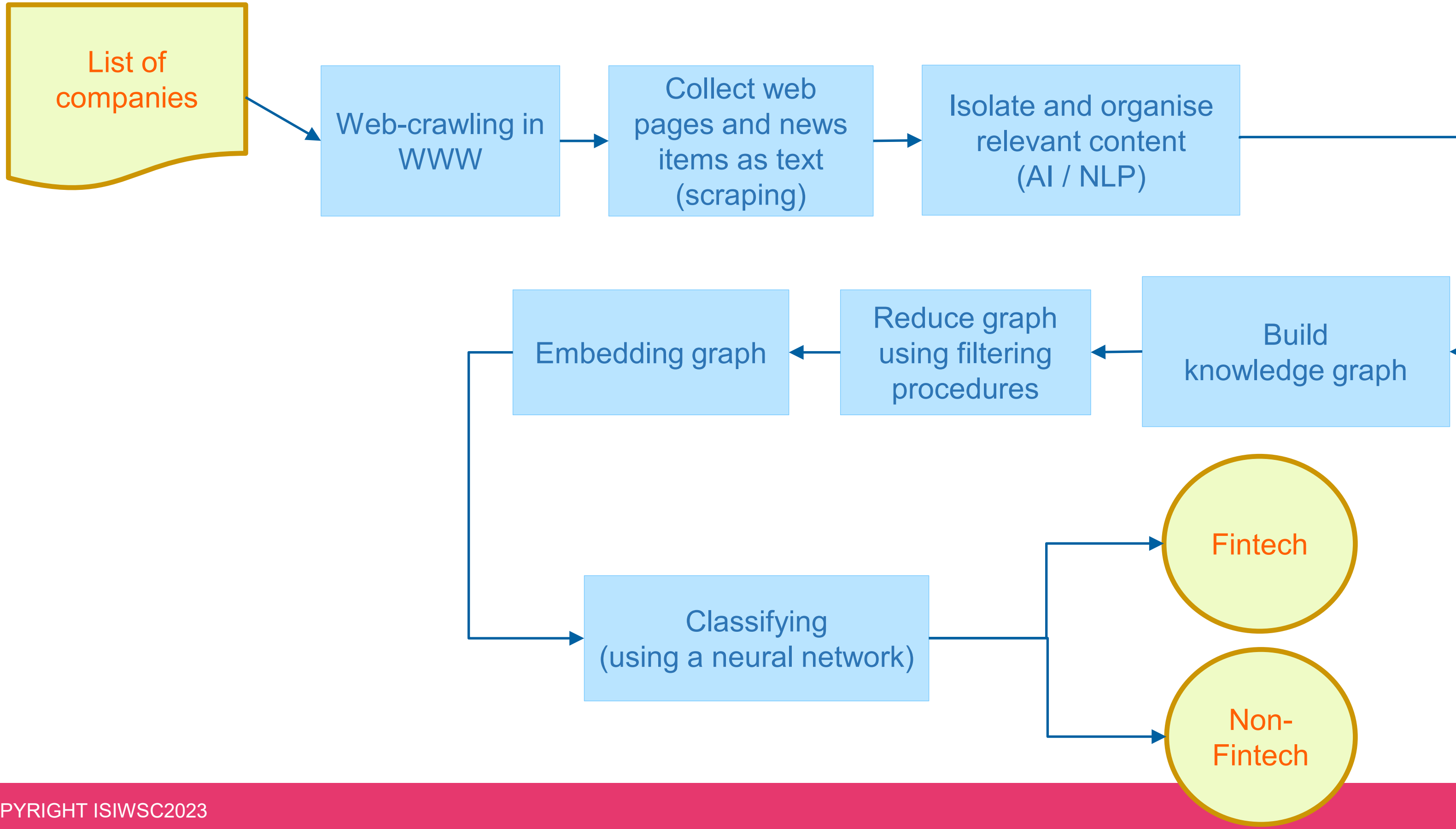
- Subset of German tech oriented companies.
- 400 fintechs
- 800 non-fintech IT companies
- No focus on optimal performance
- No claim for optimal architecture

The Knowledge Graph



- Nodes are given by **companies**, **named entities** (organisations, persons, locations) and **keywords**
- Classifying an uncategorised (new) company: is it located in the graph like a fintech or like a non-fintech?

The Knowledge Graph



List of companies

Web-crawling in WWW

Collect web pages and news items as text (scraping)

Isolate and organise relevant content (AI / NLP)

Build knowledge graph

Reduce graph using filtering procedures

Embedding graph

Classifying (using a neural network)

Fintech

Non-Fintech

Results: Interpretation and improvement potential

Summary:

- The approach as such is successful
- There is quite some potential for improvement

Data acquisition

Data organisation and graph cleaning

Binary classification

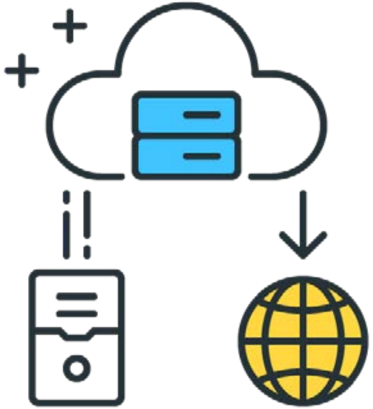
Results: Data Acquisition

Summary:

- Web scraping comes without costly licenses and time consuming tenders

- Crawling and scraping need:
  - Infrastructure services & coding
  - A reliable legal framework

- For simplicity, infrastructure services for web crawling were acquired externally.



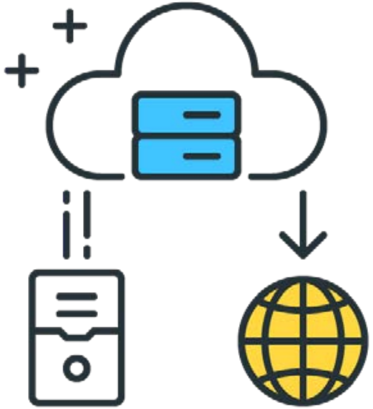
Results: Data Organisation

- AI-based extraction and organisation of contents works well

- The graph embodies the relationships between
  - *companies*,
  - *named entities (organisations, persons, locations)* and
  - *key words*

- Constructing full scale knowledge graph computationally demanding:
  - Graph (# of nodes) needs to be shrinked, using different filters

- To enable model training, the graph is embedded into a lower dimensional vector space



Results: Graph Cleaning

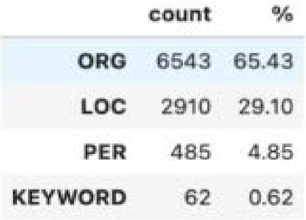
Cleaning steps:

- filtering by number of connections
- filtering by node connection distribution
- selected 10000 NEs that have the most unbalanced fraction of fintech

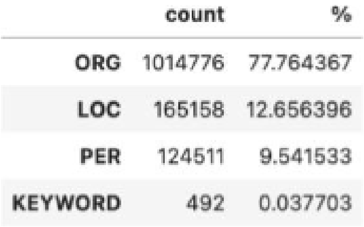
Results:

- number of nodes is decreased by factor 130
- number of edges decreased by factor 90

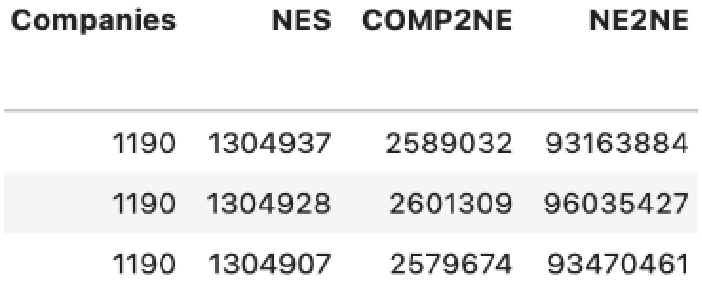
Data before and after the Graph cleaning



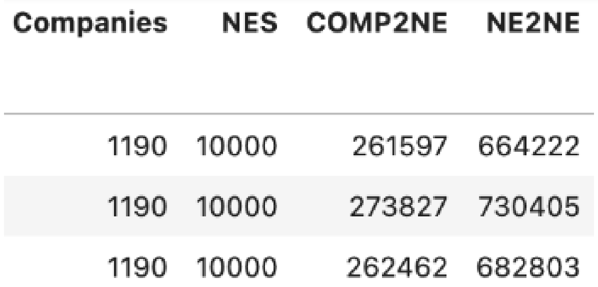
	count	%
ORG	6543	65.43
LOC	2910	29.10
PER	485	4.85
KEYWORD	62	0.62



	count	%
ORG	1014776	77.764367
LOC	165158	12.656396
PER	124511	9.541533
KEYWORD	492	0.037703



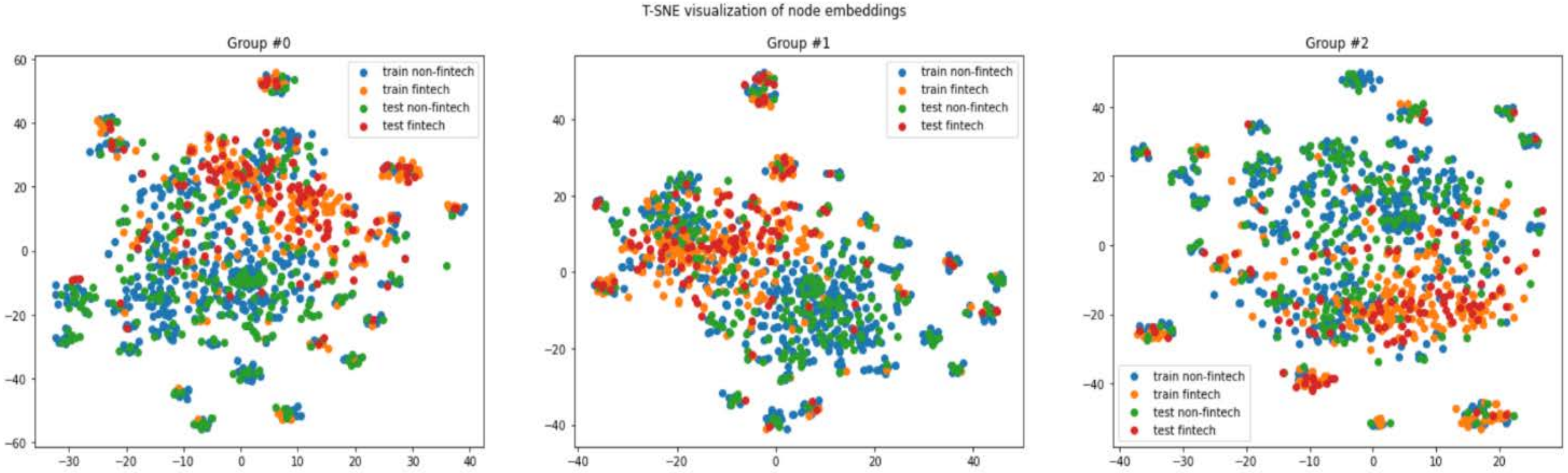
Companies	NES	COMP2NE	NE2NE
1190	1304937	2589032	93163884
1190	1304928	2601309	96035427
1190	1304907	2579674	93470461



Companies	NES	COMP2NE	NE2NE
1190	10000	261597	664222
1190	10000	273827	730405
1190	10000	262462	682803

Results: Binary Classification

Visualisation of node embedding for the three folds:



Results: Binary Classification

- PoC carried out based on a limited information set.
  - Richer training data ought to result in better performance.

- Results for baseline specification:
  - Accuracy: 88% (overall rate of correct predictions)
  - Precision: 86% (true positives as %age of selected companies)
  - Recall: 75% (true positives as %age of overall fintechs)

Bottom line:
Using a combination of public web data (news, webpages, articles) and keywords of articles, one may classify companies (as fintechs / non-fintechs) with relatively good accuracy.

Key insights

- **Legal base for web scraping** is still unstable:
  Since 2021 there are (in Germany) new provisions in copyright law allowing web scraping, but website operators may opt out, and there are no technical standards and no legal precedent as yet.
- **Buying services for subcomponents** reduces (implementation) complexity immensely, e.g. a search API service and a web content extraction service
- **Buying data services**, e.g. Pitchbook and Crunchbase, allows for more varied input data, can reduce data engineering and externalises some of the (immense) data engineering work.
- **Fintech detection is feasible**, but building a system for the related tasks (identification and later re-identification of changes, continual updating) is a **massive project**