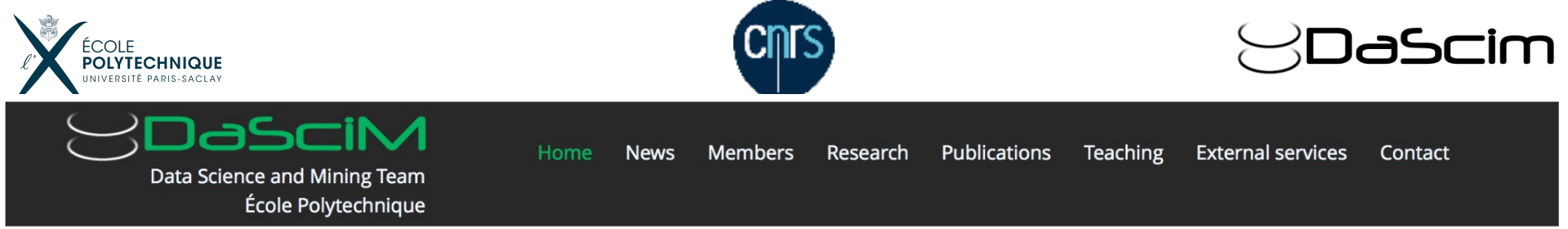


Data Science & Mining group

LIX @ Ecole Polytechnique

ANR CHAIR
Advanced Machine/Deep learning for Heterogeneous Large scale Data (ANR/HELAS)

Michalis Vazirgiannis

http://www.lix.polytechnique.fr/dascim/

September 2020

Advanced Machine/Deep learning for Heterogeneous Large scale Data (ANR/HELAS)

- Partners
  - Ecole Polytechnique – academic
  - MAIF (Insurance)
  - LINAGORA (OSS/DS)

- Time frame – 2020 -2204
- Funding for ~6 PhDs + 3 HY postdocs
- LIX team: Data Science and Mining
  - http://www.lix.polytechnique.fr/dascim/

DASCIM people

Senior Members

- C. D'Ambrosio – CR1 CNRS
- L. Liberti – DR2 CNRS
- J. Read – Assist. Prof (AXA chair)
- M. Vazirgiannis – Professor, X – Group Leader

Ph.D. students

- S. Limnios
- O. Pallanca - APHP
- G. Shang – Ph.D. – CIFRE (Linagora)
- Y. Yang – Ph.D. – CIFRE (Tradelab)
- G. Dasoulas – PhD Cifre Huawei
- G. Salha – PhD Cifre - Deezer
- M. Cerulli
- M. Kamalledine
- S. Khalife

Post doctoral researchers

- Y. Nikoletzos –post doc
- A. Tixier – post doc
- J. Lutzyer– post doc researcher

Research engineers

- C. Xypolopoulos
- K. Skianis

Interns

- Y. Guo – X
- V. Rennard – X

Machine/Deep Learning & AI, Big Data Analytics, Optimization, Text Mining/NLP, Graph Mining, Influence Max, Recommendations

Previous research highlights - NLP: Graph-based Text Mining

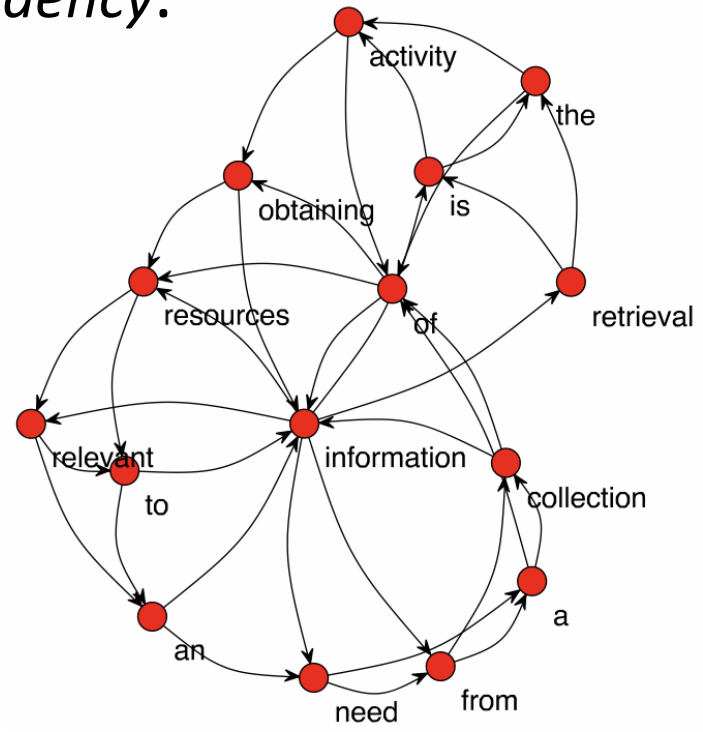
- *bag-of-words* vs. *graph-of-words:* represent a document as a *graph* to capture *word order* and *dependency*.

information retrieval is the activity of obtaining

information resources relevant to an information need

from a collection of information resources

Bag of words: ((activity,1), (collection,1)
(information,4), (relevant,1),
(resources, 2), (retrieval, 1)..)



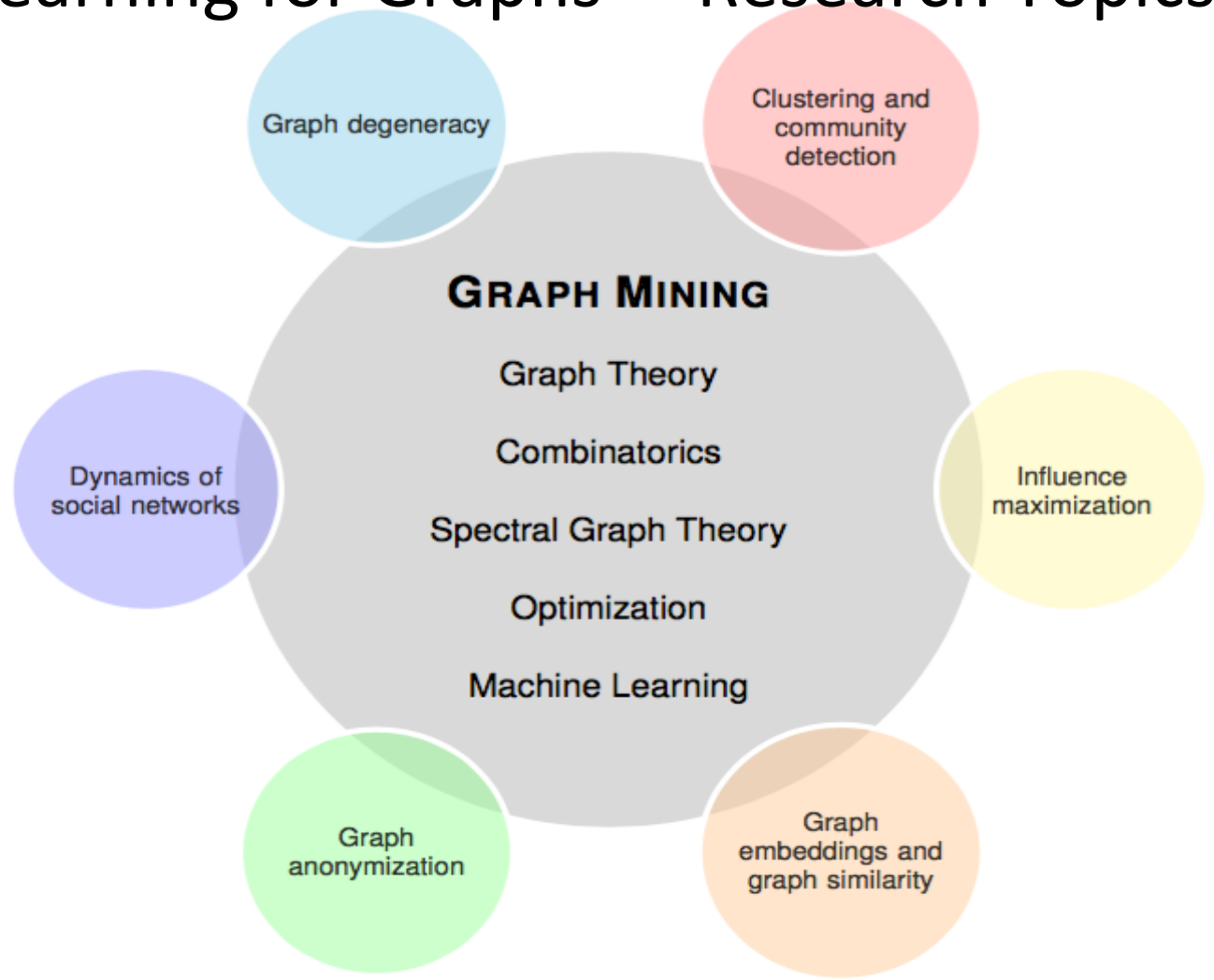
"Graph of word approach for ad-hoc information retrieval", F. Rousseau, M. Vazirgiannis, Best paper mention award ACM CIKM 2013

Previous research highlights - NLP/Text Mining - Research Contributions

Graph of Words (GoW) approach with applications to

- Ad Hoc Information Retrieval (tw-idf) [CIKM2013]
- Keyword Extraction and summarization of text streams [ECIR2015, EMNLP2016, EACL2017, ACL 2018]
- Event Detection in Textual Streams (twitter, banking,…) [ICWSM2015, ECIR2018]
- Text Categorization/opinion mining/sentiment analysis [ACL2015, EMNLP2015, EMNLP2016, EMNLP2017]
- *Document visualization and summarization* [ACL2016, ACL2018]
  - GoWis prototype software

Previous research highlights - Machine/Deep Learning for Graphs – Research Topics



Previous research highlights - Machine/Deep Learning for Graphs

- Novel metrics for node /community importance
  - Extensions of degeneracy to weighted, directed (D-core) and signed graphs [ASONAM2011, ICDM2011, KAIS2013 , SIAMDM2013]
  - Scalable Degeneracy-based graph clustering [AAAI2014]
  - 10^9 node graph clustering and community detection for fraud detection
- Identification of influential spreaders
  - Identification of influential spreaders [Scientific Reports/Nature 2016]
  - Novel influence metrics (citation and social networks) [PLOS2018][INFmetrics2019]
- Graph kernels
  - Matching Node Embeddings for Graph Similarity [AAAI 2017]
  - Degeneracy framework for graph similarity [IJCAI 2018 - best paper award]
  - Enhancing graph kernels via successive embeddings [CIKM 2018]
  - Shortest-path graph kernels for document similarity – [ENMLP 2017]
  - Grakel Python (scikit) [JMLR2020]- https://github.com>ysig>Grakel

Previous research highlights - Deep Learning for Graphs & NAS

- Kernel Graph CNN [ICANN 2018]
- Message Passing GNNs for Document Understanding [AAAI2020]
- Learning Structural Node Representations on Directed Graphs [COMPLEX NETS 2018]
- Rep the Set: Neural Networks for Learning Set Representations [AISTAS2020]
- Learning Structural Node Representations using Graph Kernels, [IEEE TKDE 2019]

ANR Chair Context

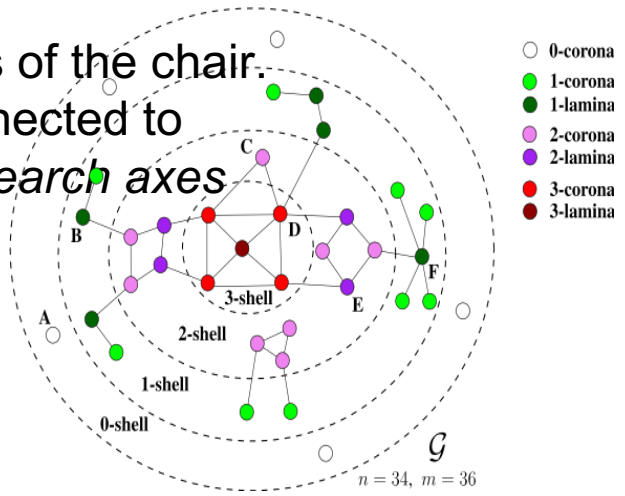
Graphs emerge a universal structure for information representation and learning for different applications

- social networks, NLP, biomedical/neuro-computing etc.

Industrial interest in graph based AI very significant

- Google having a devoted lab on relevant research
- *Tencent* organizing the Alchemy graph regression challenge
- immense scientific production in the last years in the such as *Deep Learning* for *Graph Neural Networks* and *NLP/Text Mining.*

- Both topics addressed in the chair
- Industrial partners have an explicit interest in the results of the chair.
- Research topics that are original and risky but also connected to real life applications and industrial needs. The main *research axes* are the following:



MAIN RESEARCH AXES - Learning Graph Representations

- Graph-structured data has grown in a wide range of domains (social networks, bioinformatics, chemo-informatics, NLP, pharma)
- Learning useful graph representations have many real-world applications.
- Graph Neural Networks (GNNs) have recently emerged as a general framework for addressing graph-related machine learning tasks
  - strong empirical performance
  - their expressive power and explainability
- GNNs consist of two phases:
  - *message passing phase*, nodes update their feature vector aggregating neighbours' feature vectors => vector representation for each node
  - *readout phase*, network computes a feature vector for the entire graph, applying a permutation invariant function to nodes' representations (e.g., sum, average) => aggregated to produce a graph representation.
  - GNN Limitations:
    - two-step approaches
    - simplistic nature of the permutation invariant functions - restrict the representation power of these networks.

Learning Graph Representations - Objectives

- Integration of graph kernels into the GNN design
  - In many cases, graph kernels outperform state-of-the-art GNN methods.

- Design novel neural network architectures
  - in contrast to standard GNNs, consist of a *single step.*
  - ideas from random walk graph kernels - permutation invariant,
  - Graph based feature extraction based on graph comparison
  - employed graph comparison algorithm differentiable – backprop possible
  - highly interpretable graph features extracted

- Novel approaches to *aggregate the representations of sets of nodes.*
  - new family of graph kernels - employ a message passing procedure.
  - aggregation to be more sophisticated than the traditional ones (e.g., sum, average, max).

Innovation

- envisaged GNN models; novel aggregation mechanismsl expressive node representations
- proposed architectures will be able to distinguish three fundamental graph properties which standard GNNs cannot
  (connectivity, bipartiteness και triangle-freeness)
- applications: NLP domain for insurance domain – one of the industrial partners is in this sector.

MAIN RESEARCH AXES - Deep learning methods for NLP applications / French linguistic resources

- *Deep Learning for Spoken Language Understanding and Summarization producing relevant linguistic resources for the French Language.*

- Research tasks

i. *creation of linguistic resources for French language* in cooperation with the industrial partners *Linagora* and *MAIF* (see commitment letters),

ii. Resources:

- a large corpus collected from the French Web (~330GB cleaned text)

- word embeddings (W2V and ELMO),

- n-gram data sets, stemmer etc.

iii. *Machine learning tasks*

- automated summarization, entity extraction and classification) for *insurance* (i.e. *MAIF)*

- *meeting summarization* (i.e. *Linagora*)

- *legal documents analysis* - relevant community, see: https://smartlawhub.net/people/

Chair Teaching / Training apsects

Current teaching:

- Machine/Deep Learning (M1 @ X)
- Advanced Topics in Artificial Intelligence (M1 @ X)
- Deep Learning for NLP and Graphs (M2 Data Science/MVA)
- Data Mining @ SPEIT/Jiatong, Shangai
- Machine Learning* - Tsinghua, Beijing

Professional training

- Data Science Starter Program (@EXED/X)
- Advanced AI Program (@EXED/X)

New methods will be integrated to the M2 and professional training ones

Thank You!

www.lix.polytechnique.fr/dascim/

We are hiring…