Cyber-Security: Learning from Relational Data

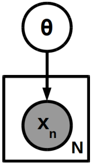
Piyush Rai

Dept. of CSE, IIT Kanpur

Nov 27, 2015

Probabilistic Machine Learning

- Machine Learning via probabilistic modeling of data



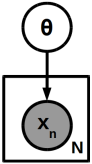
- Data assumed generated from a probability model

$$\mathbf{x}_1, \ldots, \mathbf{x}_N \sim p(\mathbf{x}|\theta)$$

- **Unified perspective**; subsume many paradigms in ML: regression, classification, clustering, dimensionality reduction, time-series analysis, etc.

Probabilistic Machine Learning

- Machine Learning via probabilistic modeling of data



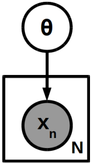
- Data assumed generated from a probability model

$$\mathbf{x}_1, \ldots, \mathbf{x}_N \sim p(\mathbf{x}|\theta)$$

- **Unified perspective**; subsume many paradigms in ML: regression, classification, clustering, dimensionality reduction, time-series analysis, etc.
- Handling **missing data, outliers, skewness, temporal evolution**, etc.
- Easily incorporate prior beliefs, quantify uncertainty **(Bayesian learning)**

Probabilistic Machine Learning

- Machine Learning via probabilistic modeling of data



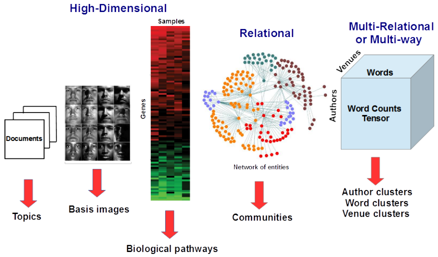
- Data assumed generated from a probability model

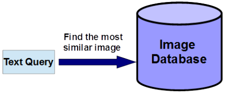
$$\mathbf{x}_1, \ldots, \mathbf{x}_N \sim p(\mathbf{x}|\theta)$$

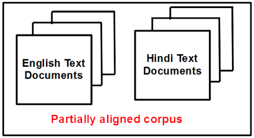
- **Unified perspective**; subsume many paradigms in ML: regression, classification, clustering, dimensionality reduction, time-series analysis, etc.
- Handling **missing data, outliers, skewness, temporal evolution**, etc.
- Easily incorporate prior beliefs, quantify uncertainty **(Bayesian learning)**
- Enables principled combinations of simpler paradigms to solve complex tasks

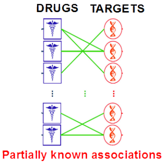
My Research

- Learning from complex, messy, heterogeneous data
- Focus: Feature representation learning, Cross-modal/Transfer Learning, reducing labeling costs, scalability (online/distributed learning), model-size adaptability as data grows (nonparametric models)









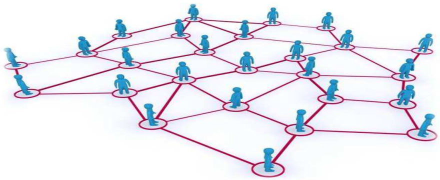
Relational Data in the "Networked" World

- A natural represention of various types of data in the cyber-space, e.g.,
  - Physical networks
  - Links between webpages
  - Links between users on social networks
  - Links across networks
- Almost all of this data is also **time-evolving** in nature

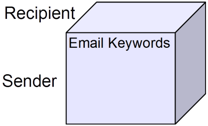
Picture source: growingsocialmedia.com

Relational Data as Graphs

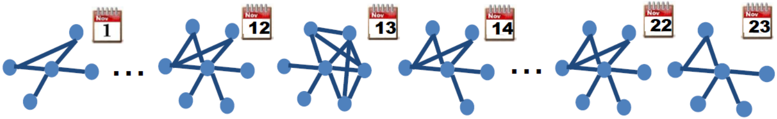
- Unipartite graphs and bipartite graphs



- Multi-dimensional graphs ("tensors")



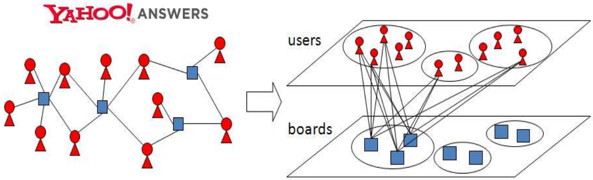
- Dynamic/time-evolving graphs



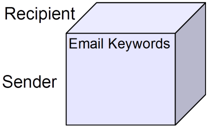
Picture source: www.aboutdm.com

Relational Data as Graphs

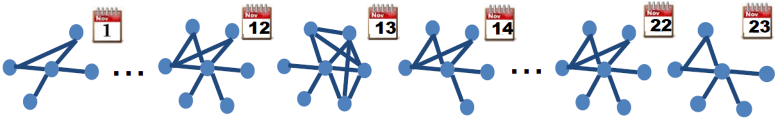
- Unipartite graphs and bipartite graphs



- Multi-dimensional graphs ("tensors")



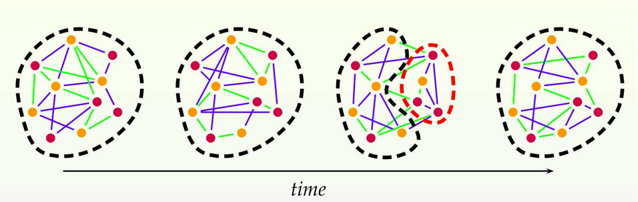
- Dynamic/time-evolving graphs



Picture source: www.aboutdm.com

Cyber-Security Problems on Relational Data

- Anomaly detection
  - Anomalous nodes
  - Anomalous edges
  - Anomalous sub-communities
- Detecting/predicting change-points in evolving graphs

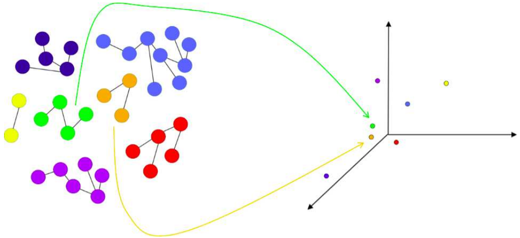


- Learning good representations of the data is key

Picture courtesy: http://www.cis.jhu.edu/ parky/

Embeddings to Represent Relational Data

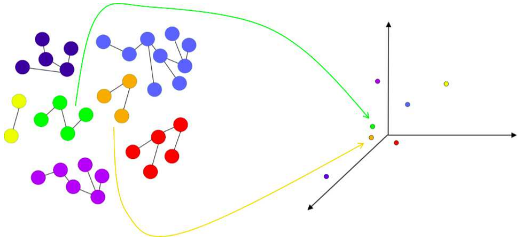
- Embedding relational data (graphs) to a vector space



- Allows performing various machine learning tasks such as clustering, classification, anomaly detection, etc., using the learned embeddings
- A challenging problem in general

Embeddings to Represent Relational Data

- Embedding relational data (graphs) to a vector space

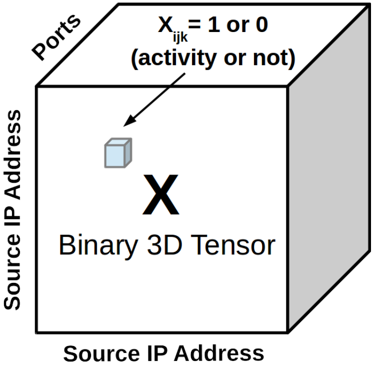


- Allows performing various machine learning tasks such as clustering, classification, anomaly detection, etc., using the learned embeddings
- A challenging problem in general

[ICDM 2013, ICML 2014, ECML 2015, UAI 2015, NIPS 2015 + forthcoming]

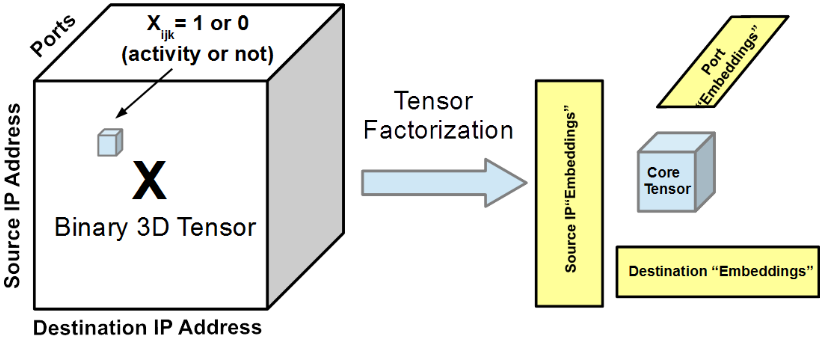
Tensors or "Multi-Relational" Data

Tensors are useful for encoding multi-way relations



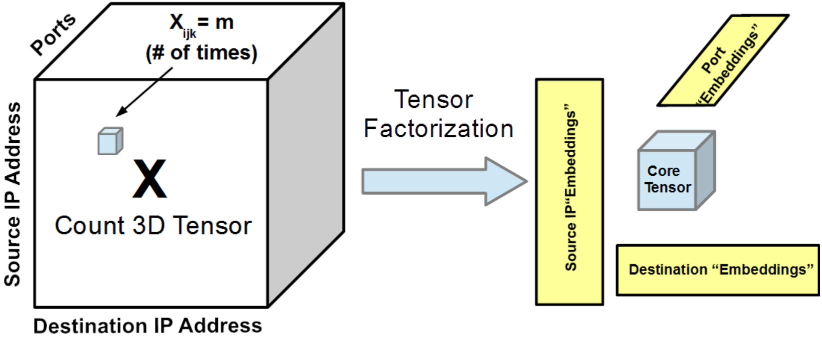
Tensor Factorization

Tensor factorization allows embeddings the entities of each "way" of the tensor



Tensor Factorization

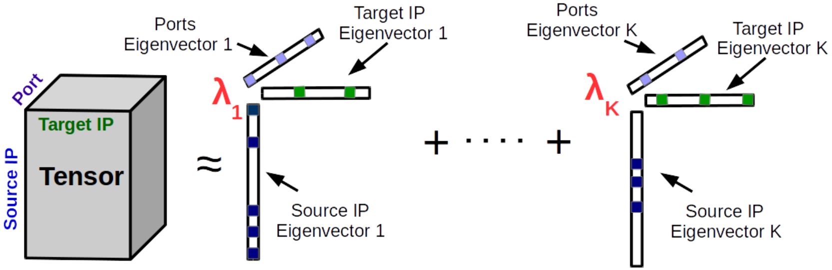
Tensor factorization allows embeddings the entities of each "way" of the tensor



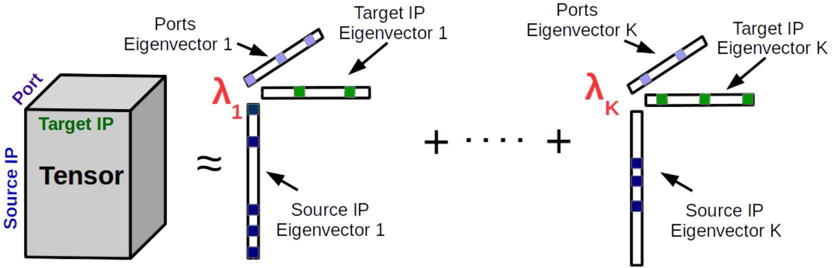
Embeddings can be used as a feature representation

Tensor Factorization

An SVD like approach: express a tensor as a weighted sum of K rank-one tensors

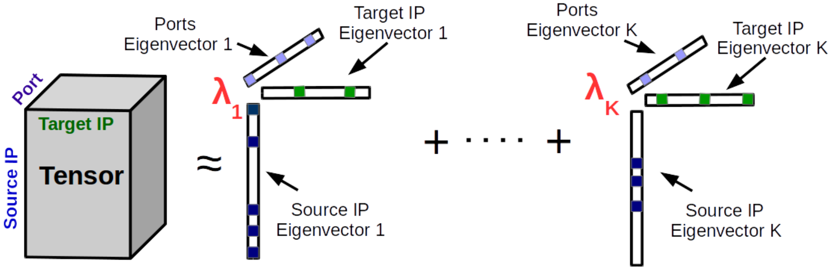


Tensor Factorization: Challenges



- Significant amounts of of missing data
- Interpretability (desirable, e.g., sparse and non-negative eigenvectors)
- Modeling the time-dimension (useful for forecasting)
- Integrating sources of side-information
- Scalability for massive tensors

Tensor Factorization: Challenges



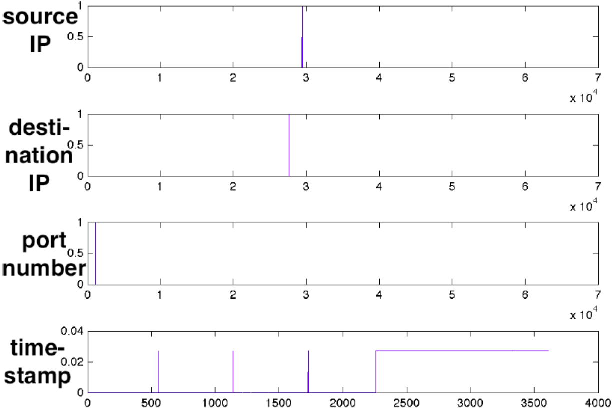
- Significant amounts of of missing data
- Interpretability (desirable, e.g., sparse and non-negative eigenvectors)
- Modeling the time-dimension (useful for forecasting)
- Integrating sources of side-information
- Scalability for massive tensors

[ICML 2014, AAAI 2015, ECML 2015, UAI 2015 + forthcoming]

Tensor Factorization for Anomaly Detection

Data: a four-way tensor Source IP $\times$ Dest. IP $\times$ Port $\times$ Time

Eigenvectors of this tensor can identify normal vs anomalous behavior[1]

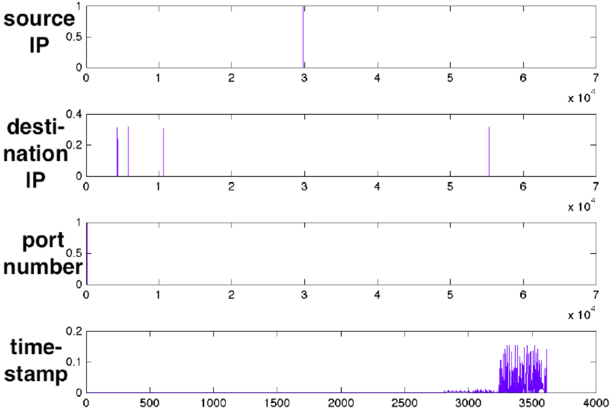


[1] Koutra *et al.* (2012)

Tensor Factorization for Anomaly Detection

Data: a four-way tensor Source IP $\times$ Dest. IP $\times$ Port $\times$ Time

Eigenvectors of this tensor can identify normal vs anomalous behavior[1]



[1] Koutra *et al.* (2012)

Other Work Relevant to Cyber-Security

- Probabilistic modeling of rare-events
- Learning only from positive examples
- Time-series of count-data (via Poisson Processes)
- Privacy preserving machine learning (private Bayesian inference)

Thanks! Questions?