Predicting Unseen Links Using Learning-based Matrix Completion

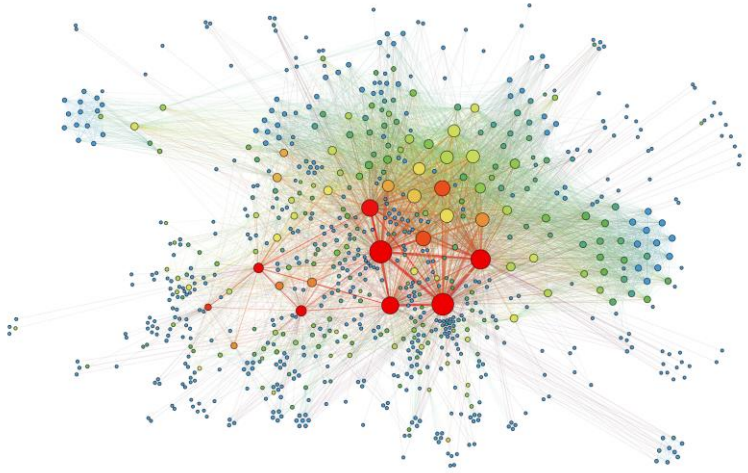
NOMS 2022

Shuying Zhuang, Jessie Hui Wang, Jilong Wang, Changqing An, Yuedong Xu, Tianhao Wu

Tsinghua University

AS-level topology is important

- AS-level topology: use nodes to represent ASes, and use edges to represent AS links



AS

AS link

- **An accurate and complete topology** is beneficial for understanding, operating, and diagnosing the Internet
  - Protocol design, performance analysis, fault diagnosis, security analysis
  - ……

AS-level topology is far from complete

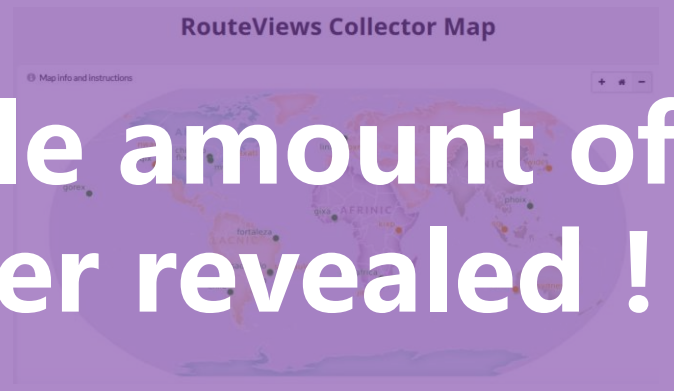
- Due to the distributed nature of the Internet, such a complete AS-level topology is not readily available !

- Significant efforts have been spent in deploying VPs
  - The number and coverage of the **VPs are still limited**
  - The AS-level topology observed from current measurement data is still **far from complete**

There remains a considerable amount of unseen AS links to be further revealed !

CAIDA Ark VPs (100+)

RIPE RIS VPs (in the upper tier)



RouteViews VPs (in the upper tier)

How to improve topology completeness

- Opportunistically conducted measurements
  - Deploy more VPs and probe more source-destination pairs
  - Neither carefully select VP locations nor select source-destination pairs

Some candidate VP locations may provide a redundant view

Resources of VPs may be limited

✗ Inefficient ! Consume a lot of resources but not bring equivalent improvements

How to improve topology completeness

- Opportunistically conducted measurements
  - Deploy more VPs and probe more source-destination pairs
  - Neither carefully select VP locations nor select source-destination pairs

Some candidate VP locations may provide a redundant view

Resources of VPs may be limited

✗ **Inefficient ! Consume a lot of resources but not bring equivalent improvements**

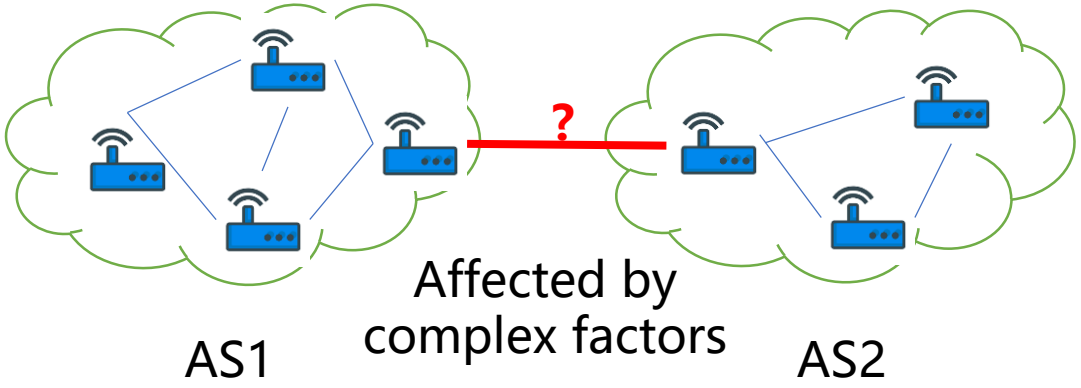
- Targeted measurements
  - Under the guidance of specific data sources

✗ **Only discover specific types of unseen links due to the limitation of the data source it uses** (e.g. IXP data)

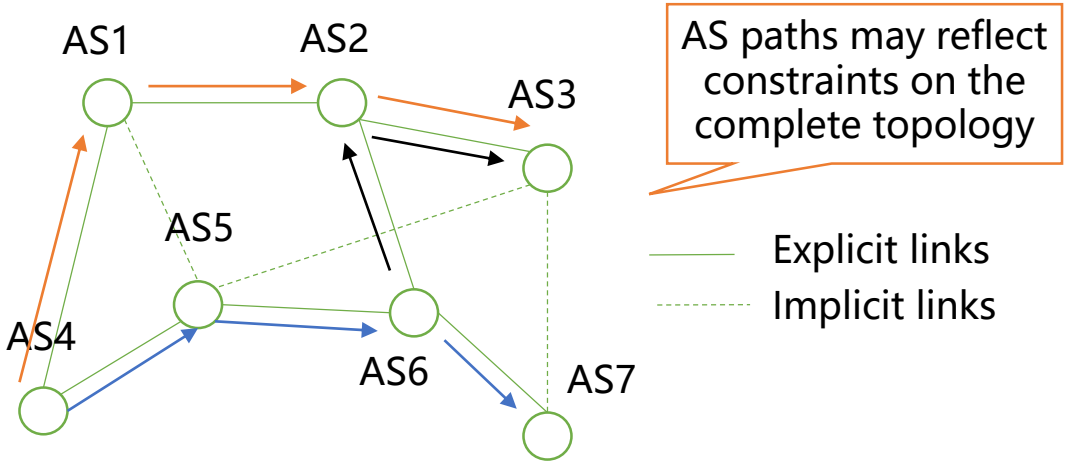
How to efficiently discover more unseen Links

Our goal: predict general unseen AS links

- **Use multi-source data** to help predict **general unseen AS links** for further efficiently discover them



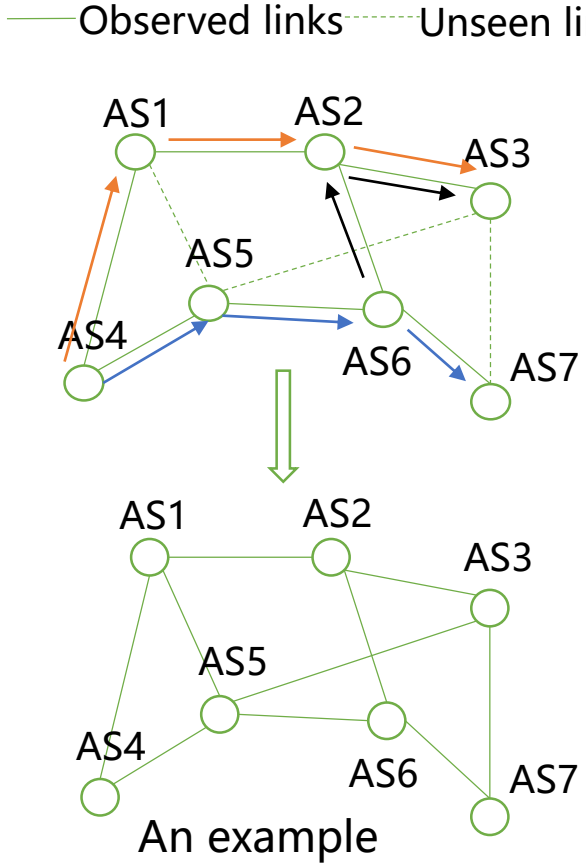
Making the prediction task challenging



By integrating such constraints, it is possible to obtain some implicit connectivity informatio

Problem formulation

- Our problem is to predict unseen AS links given a set of observed AS paths

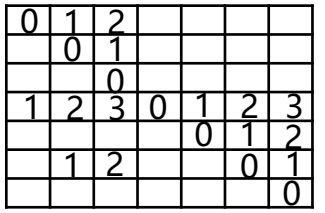


Define an **AS hop count matrix H** to hold the observed AS paths. H contains a small number of known entries and many unknown entries.

Key issue

H should be **low-rank**
AS hop count matrix completion: leverage known entries to recover all unknown entries

Predicting unseen links by comparing the recovered hot counts with a threshold T



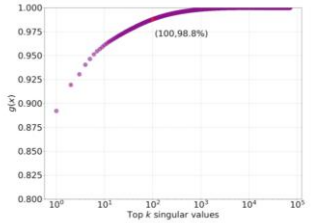
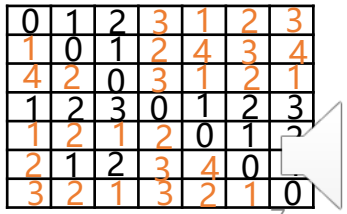


Fig. 1. The percentage of the total variance that can be captured by the top k singular values.



Our AS side-information assisted GMF-based completion method (SGMF)

- Develop a learning-based matrix completion method specifically for the unseen AS link prediction problem
  - Traditional completion methods have limited learning capabilities
  - Some learning-based completion methods lack valuable side-information

Inherit the **powerful learning capability** of neural networks in GMF

Integrate **valuable side-information** based on the understanding of peering practices

Learn more expressive AS latent vectors and achieve outstanding prediction performance

Our AS side-information assisted GMF-based completion method (SGMF)

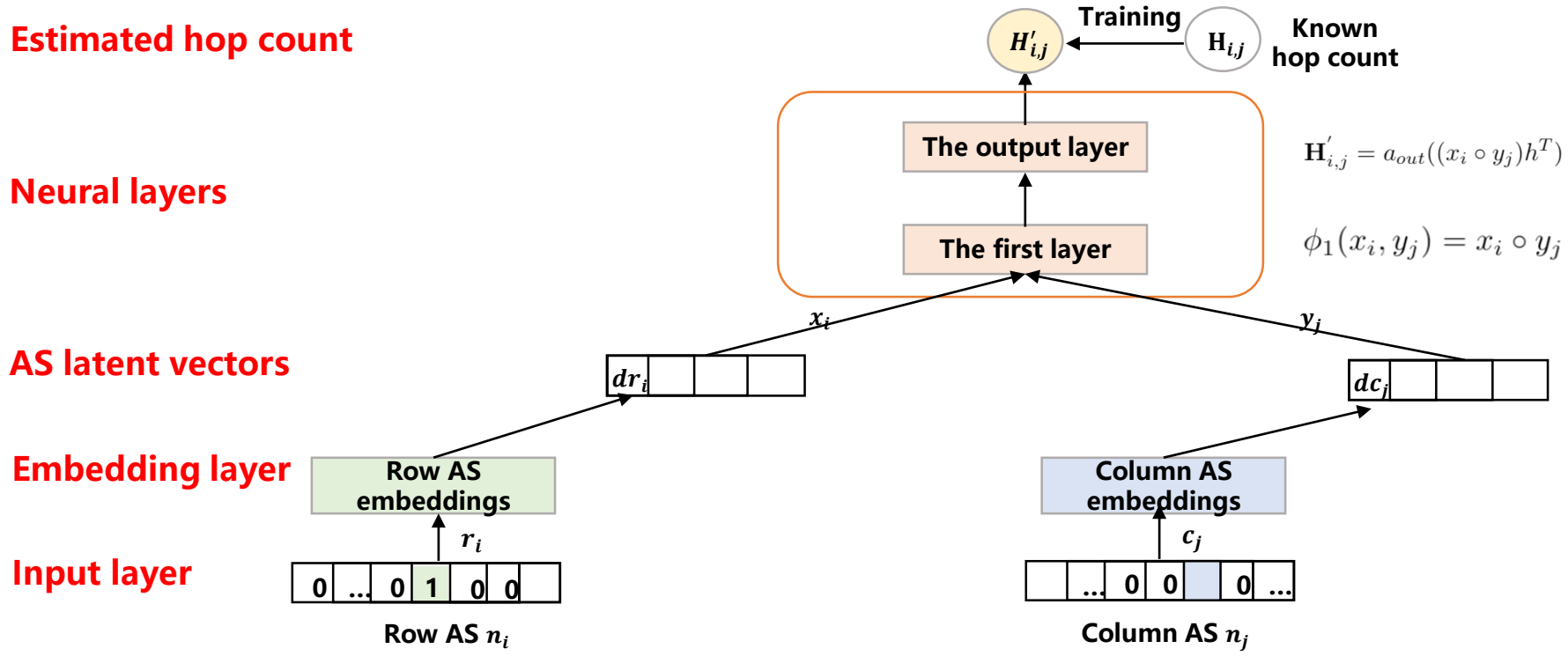
- There are multi-source information related to the inter-domain topology and routing
- AS side-information:

AS tier feature	AS tier feature
Traffic information features	overall traffic volume feature、traffic ratio feature
Peering policy features	the general policy feature, multiple locations requirement feature, traffic ratio requirement feature and contract requirement feature
Geographic scope feature	Geographic scope feature
AS type feature	AS type feature

- **Valuable for recovering unknown AS hop counts**

Our AS side-information assisted GMF-based completion method (SGMF)

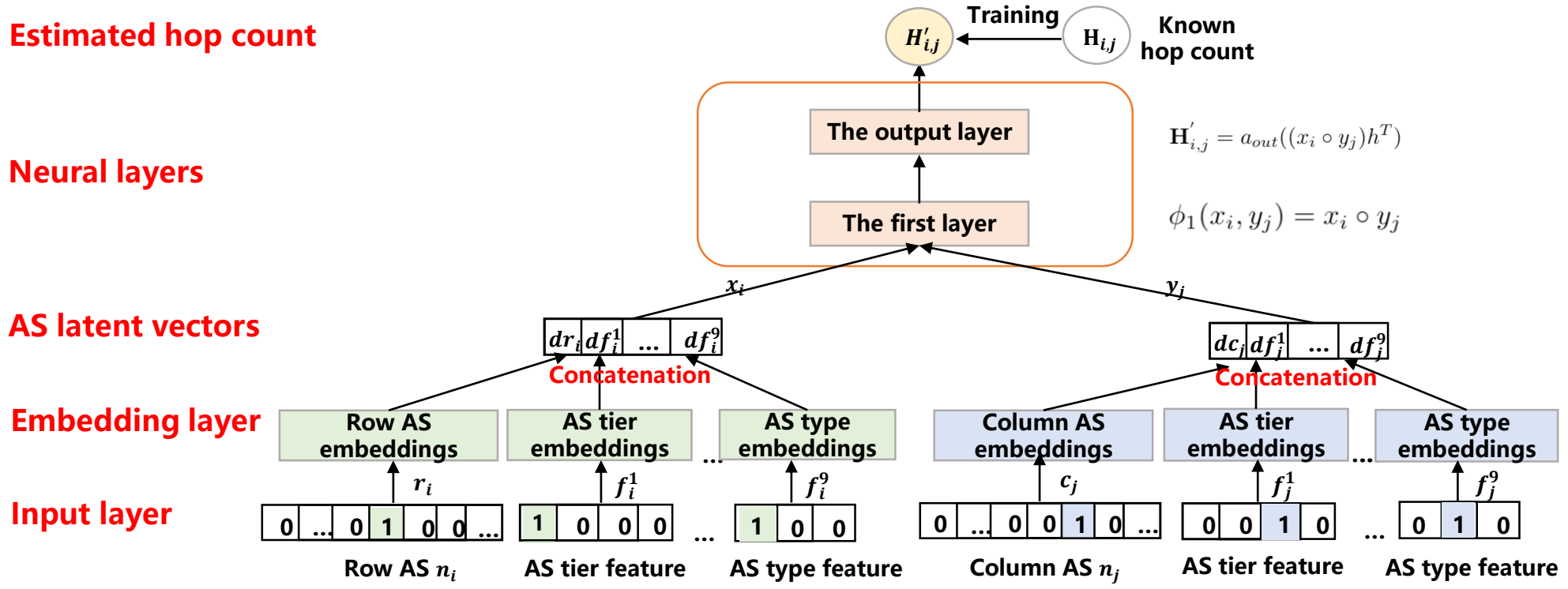
- Framework Design



- Model parameters (e.g. embeddings and h) can be learned from known AS hop counts

Our AS side-information assisted GMF-based completion method (SGMF)

- Framework Design



- Model parameters (e.g. embeddings and h) can be learned from known AS hop counts

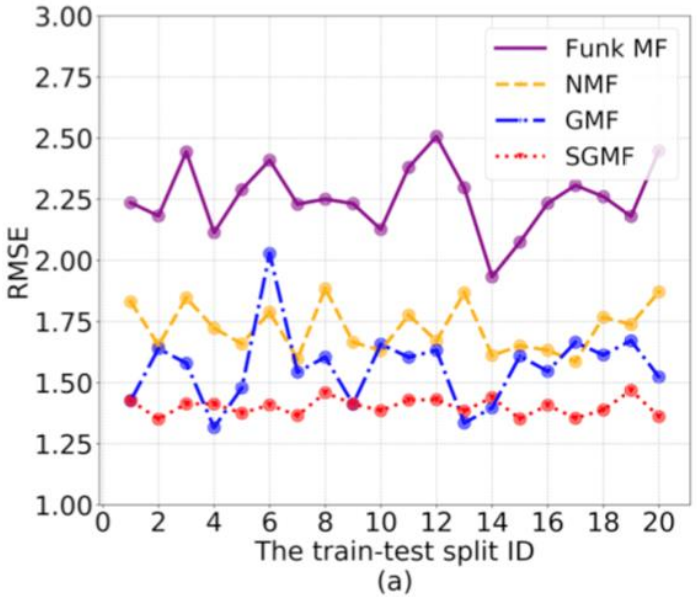
Performance evaluation

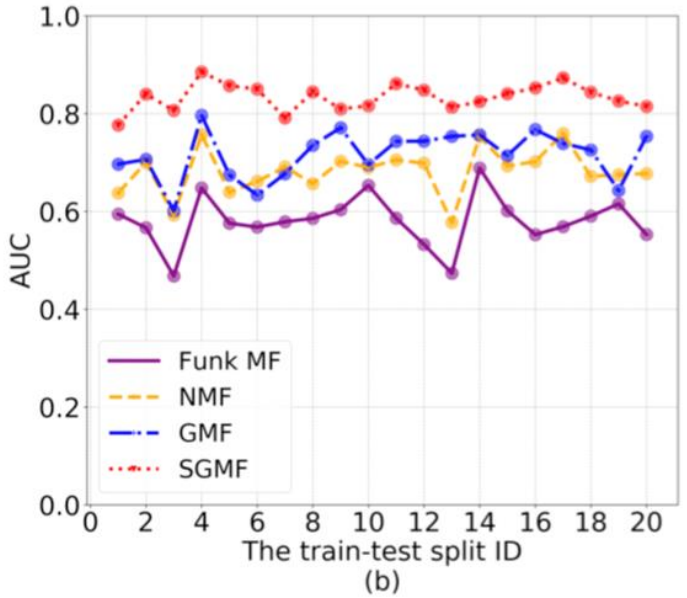
- Datasets
  - **AS paths** from one BGP table observed by 1134 VPs in RouteViews and RIPE RIS
  - Collect most **side-information** from PeeringDB
- Evaluation Metrics
  - Root Mean Squared Error (RMSE)
  - The area under the receiver operating characteristic curve (AUC)
  - True positive rate (TPR) and false positive rate (FPR)
- Evaluation Methodology
  - AS paths observed from 85% (15%) VPs for training (testing)
  - Create **20 random train-test splits** for cross-validation

Comparison with other completion methods

- Funk matrix factorization method (Funk MF)
- Nonnegative matrix factorization method (NMF)
- GMF

SGMF **improves the AS hop count estimation performance** by **10.3%** even compared with GMF





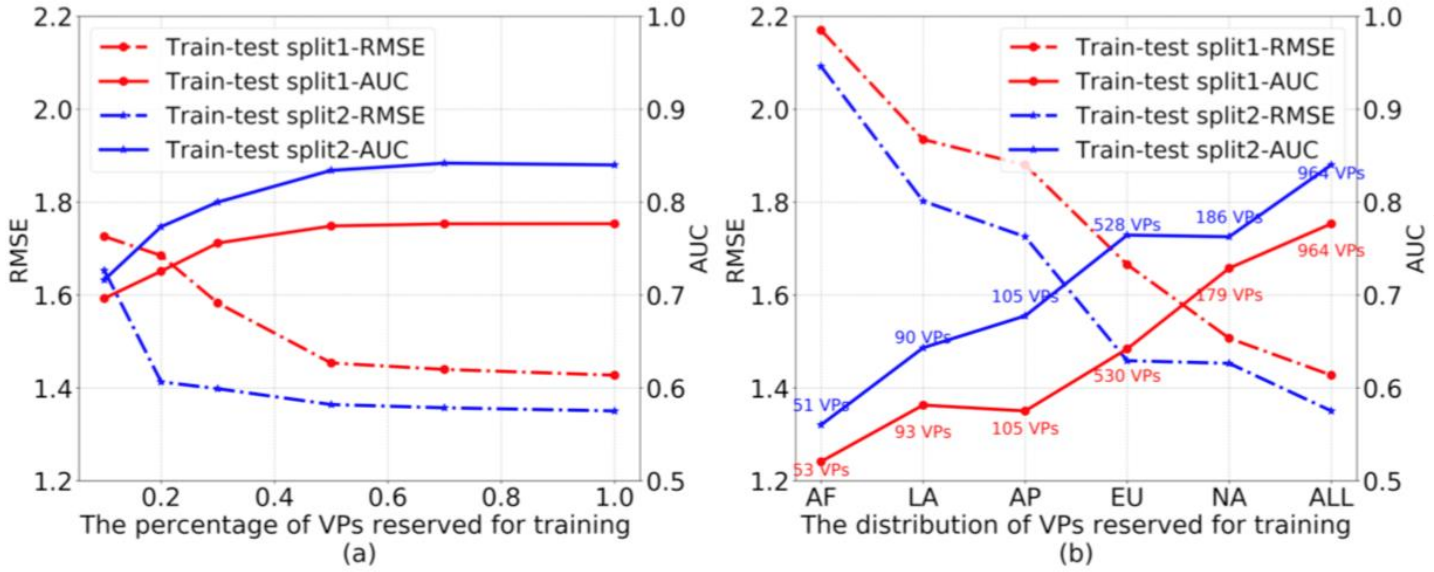
SGMF **improves the link prediction performance** by about **17% to 44%** than other completion methods

Fig. 3. (a) RMSE of Funk MF, NMF, GMF and SGMF on different train-test splits; (b) AUC of Funk MF, NMF, GMF, SGMF on different train-test splits.

Impact of the number and distribution of known measurement VPs

- Randomly select 10%, 20%, 30%, 50%, 70% of VPs
- Select VPs in each region
  - Africa (AF), Asia-Pacific (AP), Europe (EU), Latin America (LA), and North America (NA)

Randomly deploying more VPs may bring a lot of redundant AS hop count information



Biased distribution of VPs may bring much biased AS hop count information

Fig. 4. (a) The RMSE and AUC performance of SGMF with respect to different percentages of VPs reserved for training; (b) The RMSE and AUC performance of SGMF with respect to different distributions of VPs reserved for training.

Unseen link prediction results

- We choose T to be 2.94 to achieve a TPR of 70.09% and a low FPR of 12.94%
  - **Bring significant topology improvements** of observing 70.09% unseen links
  - **Save resources of launching unnecessary measurements** for finding 87.06% (1-FPR) non-links

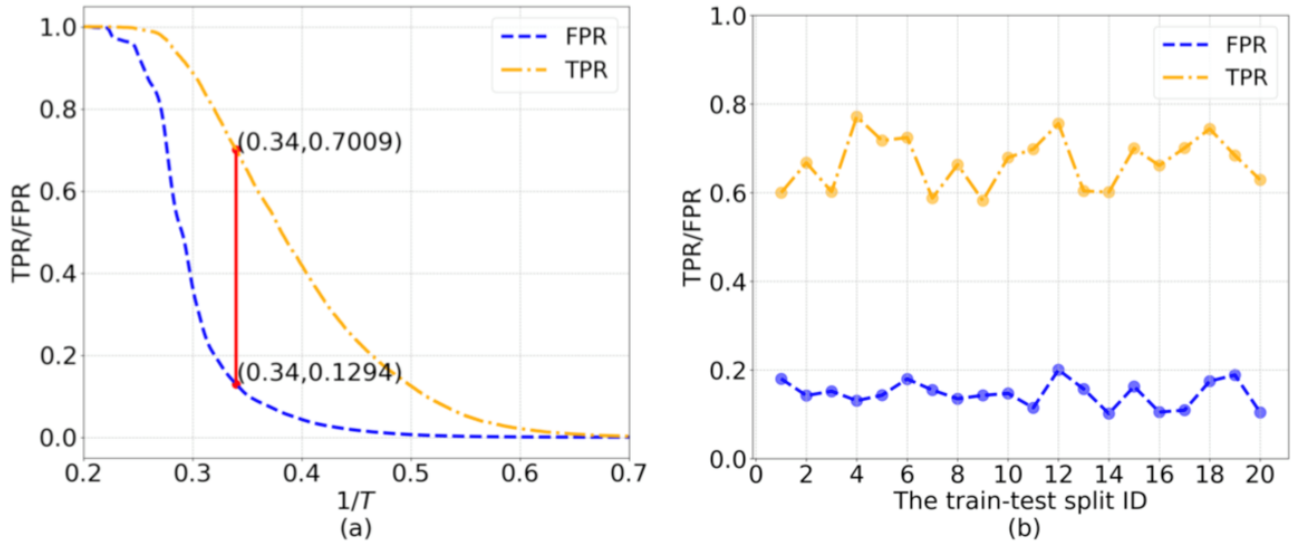


Fig. 5. (a) The FPR and TPR of the trained SGMF model under different T; (b) The distribution of FPR and TPR of trained SGMF models under other different train-test splits with T=2.94.

Validation using different datasets

- CAIDA Ark’ s IPv4 Routed /24 AS Links Dataset
- AS paths returned by Looking Glass VPs

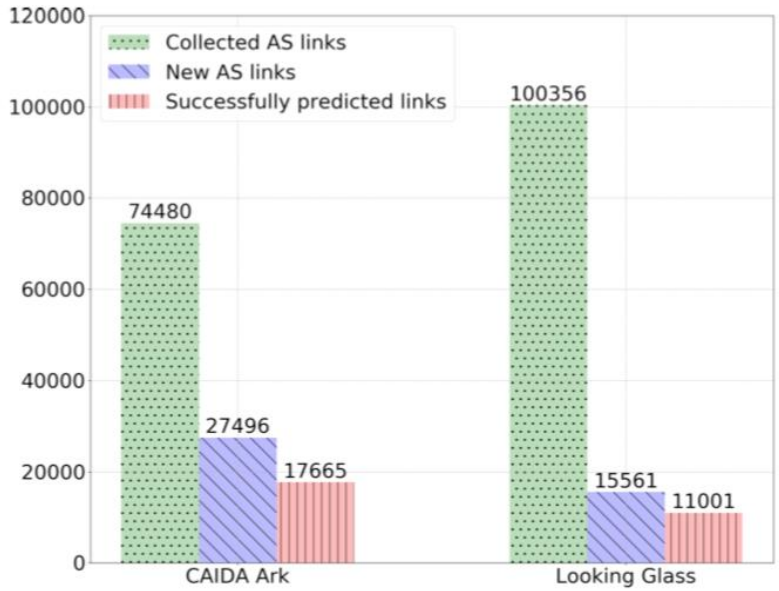


Fig. 6. The number of observed AS links, new AS links and successfully predicted AS links from each dataset.

- ✓ New AS links actually verify the existence of a large number of predicted AS links
- ✓ About 65% of the new links have been correctly predicted

Significant completeness improvements can be achieved under the guidance of the predicted AS links

Conclusion

- Develop **a learning-based matrix completion method SGMF** to predict Internet-scale unseen links
  - Exploits a neural network and utilizes side-information
  - Improve AS link prediction performance by 17%
  - Bring topology improvements of observing 70.09% unseen links
  - Save resources of launching unnecessary measurements for finding 87.06% non-links

- Ongoing work
  - Conduct measurements under the guidance of our predicted links to discover them efficiently

Thanks for your attention!
Q & A

Contact:
zhuangsy18@mails.tsinghua.edu.cn

清華大學
Tsinghua University