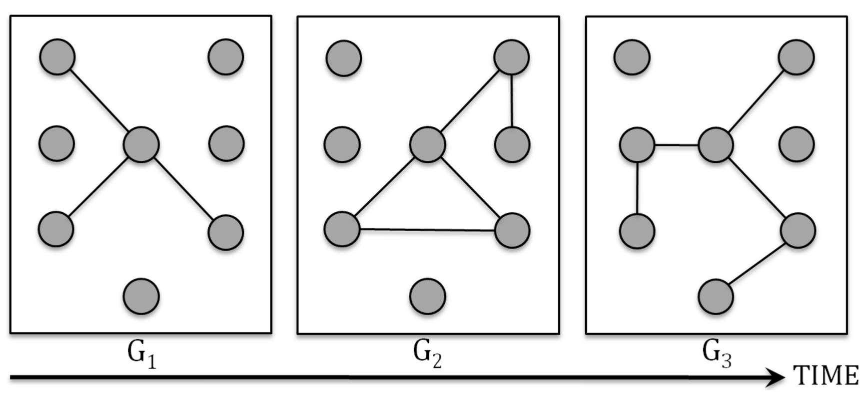
# Modeling the Evolution of Discussion Topics and Communication to Improve Relational Classification

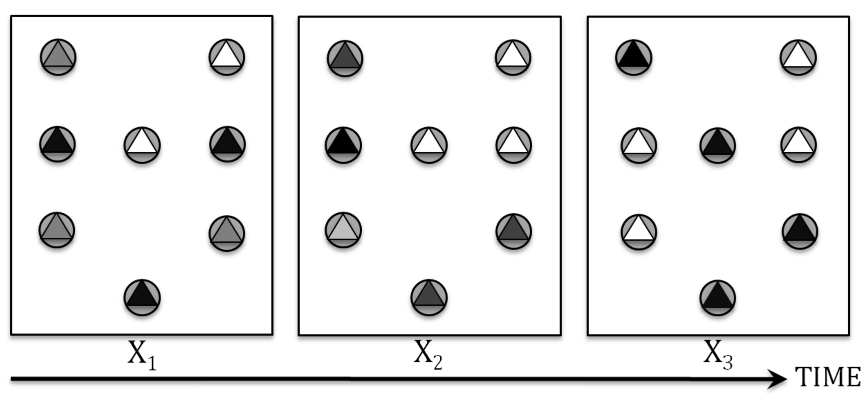
Ryan Rossi and Jennifer Neville, CS Department, Purdue University

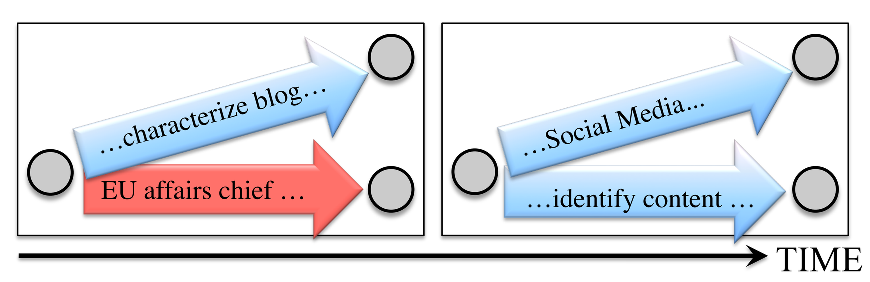
PURDUE UNIVERSITY

## Introduction

- Although relational dependencies have been successfully exploited in classification models, most approaches ignore temporal network information and only consider *static* network snapshots
- However, many relational domains have both network structure and attributes changing over time
  - For example, in social media there can be temporal dynamics in both the communication structure and message/document content
- We aim to exploit these dependencies between temporal and relational information to improve predictive accuracy
- **Key ideas**:
  - Events in the recent past are more influential than events in distant past
  - Regular series of events are likely to indicate stronger relationships than events isolated in time

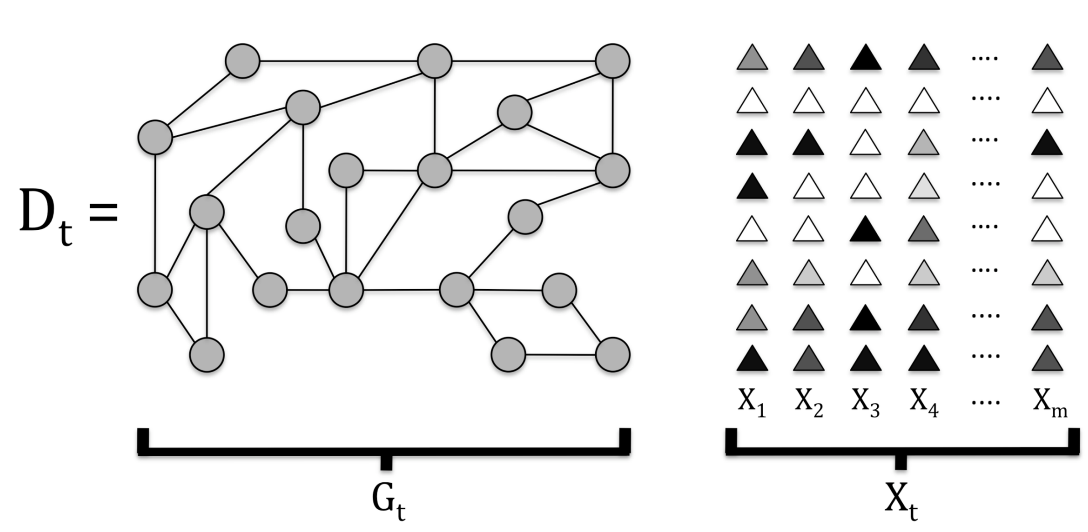


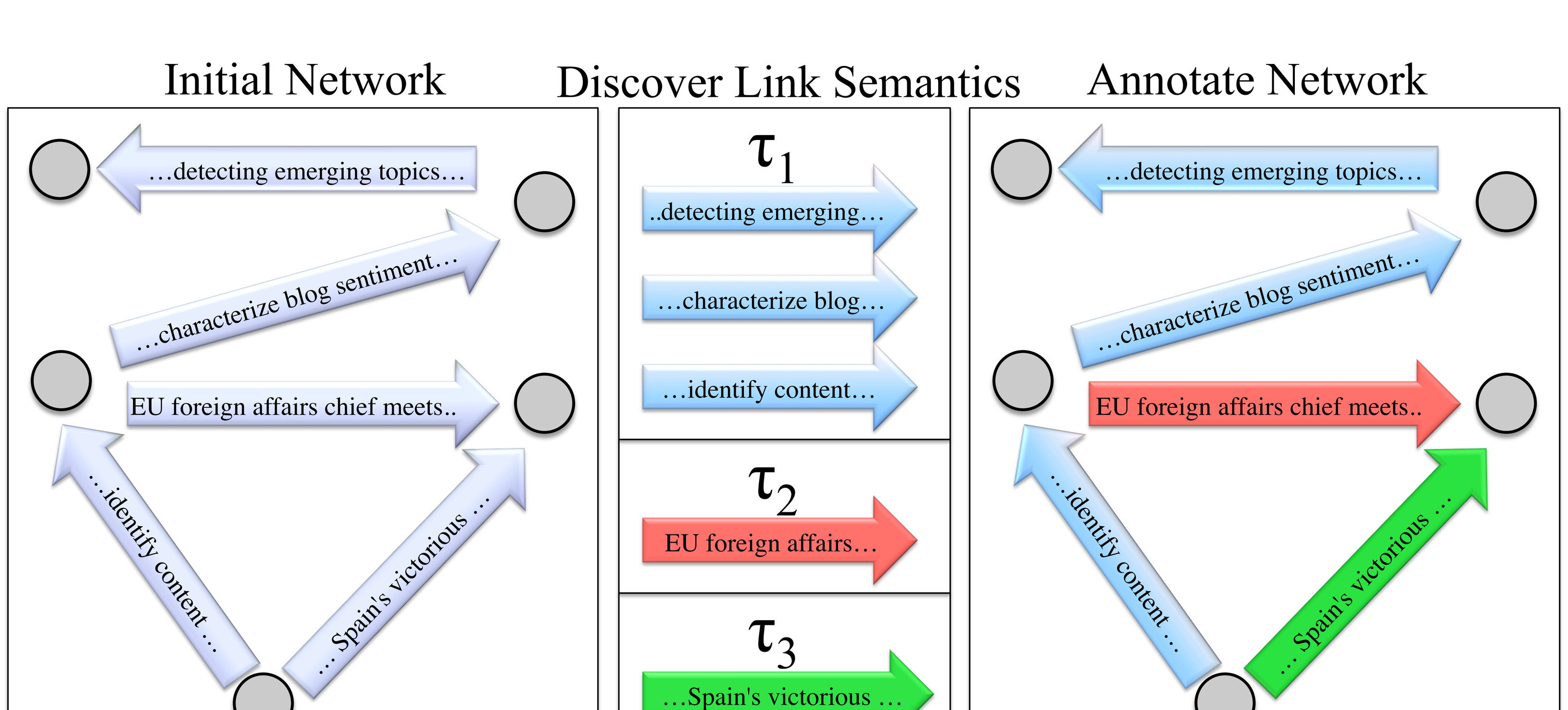

## Data: Python Open Source Development

- We extracted emails and bug discussions from the open-source python development environment (01/01/07 - 09/30/08)
  - 13181 email messages from 1914 developers
  - 69435 bug comments from 5108 developers
- Let D = $D_1, D_2, \dots, D_n$ be a sequence of temporal snapshots
  - Every temporal snapshot *i* corresponds to the events that occurred during the time period *i*.
  - The size of the temporal snapshots are three month periods.
- **Goal**: Predict individual developer effectiveness (*has closed bug*) given the communications between developers and their latent topics.




## Textual Analysis: Interpreting Links and Nodes

- Initial dataset has only developer emails and bug discussions
- **Network Annotation:** Automatically annotate the links and nodes by discovering the latent topics of the communications between individuals
- Motivation:
  - In the task of predicting effectiveness we may find that communications about specific topics may indicate more *productive* interactions
  - For example, communications about 'sports' may correspond to less effective interactions than those discussing 'web programming'
- We have developed a simple method for assigning such semantics to the links and nodes in a text-based network.
  - Use LDA to identify communication topics
  - Label each communication link with it's most likely latent topic and each individual with their most frequent topic of communication.




## Temporally-Evolving Network Classifier

- **Phase 1: Model Temporal Influence of Links and Attributes**
  - Transform dynamic graph into statically weighted summary graph and set of weighted summary attributes using kernel smoothing (exponential kernel)

**Attribute Summarization**

$$\mathbf{X}_{S_t}^V = \mathbf{X}_1^V \cup \mathbf{X}_2^V \cup \cdots \cup \mathbf{X}_t^V$$

$$\mathbf{X}_{S_t}^E = \mathbf{X}_1^E \cup \mathbf{X}_2^E \cup \cdots \cup \mathbf{X}_t^E$$

$$K_X(\mathbf{X}_i; t, \lambda) = (1-\lambda)^{t-i} \lambda W_i^X$$

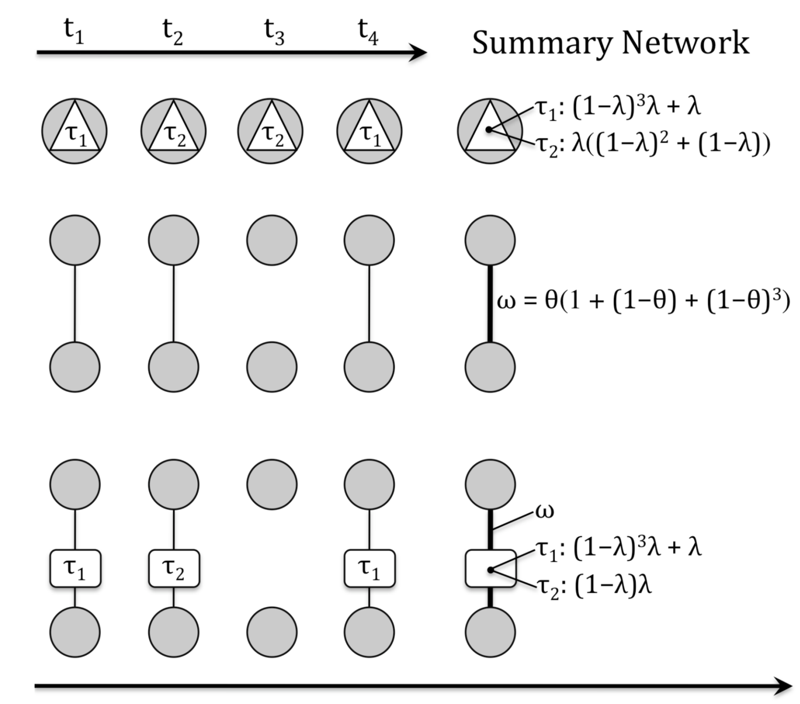
$$W_{S_t}^X = \beta_1 W_1^X + \beta_2 W_2^X + \cdots + \beta_t W_t^X = \sum_{i=1}^{t} K_X(\mathbf{X}_i; t, \lambda)$$

**Graph Summarization**

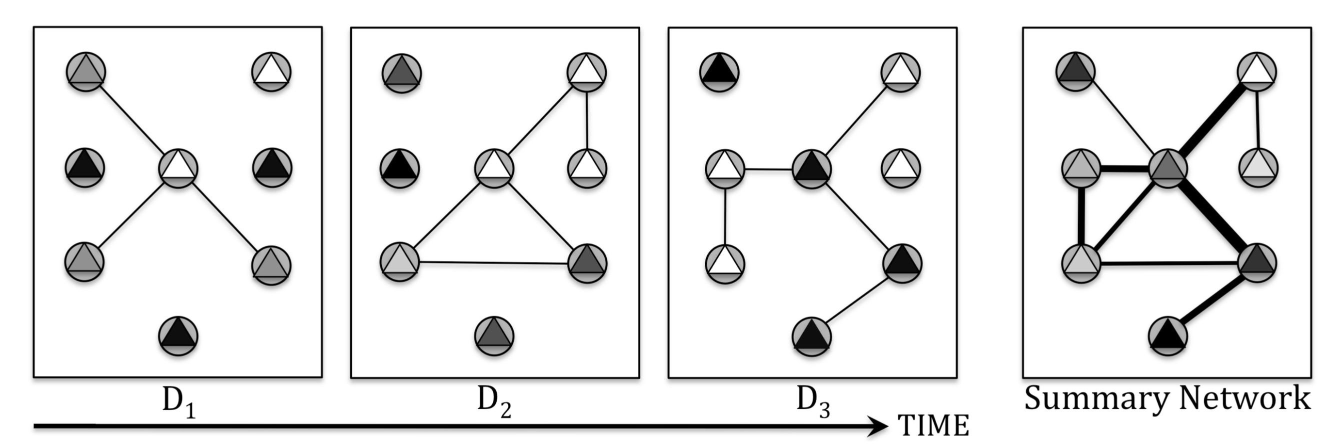
$$E_{S_t} = E_1 \cup E_2 \cup \cdots \cup E_t$$

$$K_E(G_i; t, \theta) = (1-\theta)^{t-i} \theta W_i^E$$

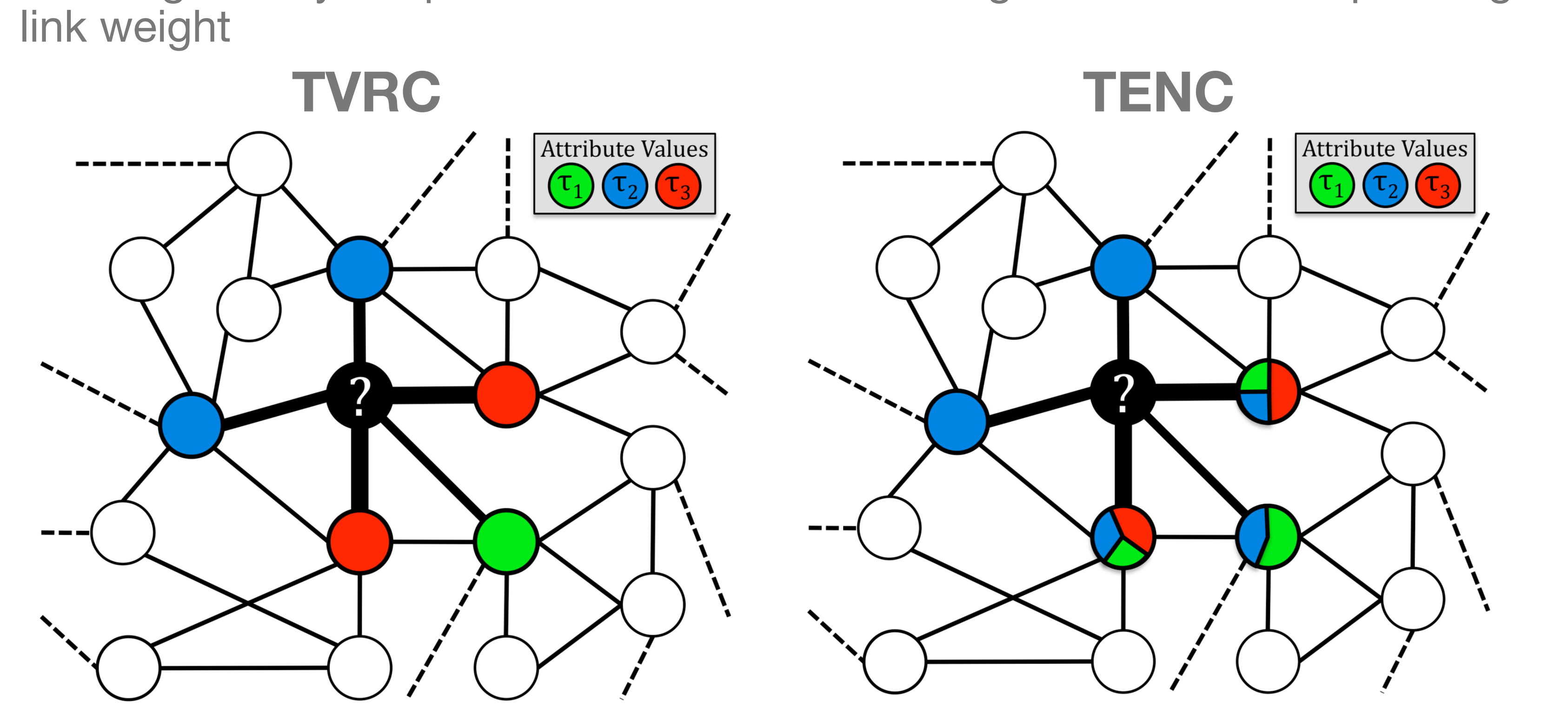
$$W_{S_t}^E = \alpha_1 W_1^E + \alpha_2 W_2^E + \cdots + \alpha_t W_t^E = \sum_{i=1}^{t} K_E(G_i; t, \theta)$$




  - Weights can be viewed as probabilities that a relationship (or attribute value) is still active at the current time step *t*, given that it was observed at time *(t-k)*



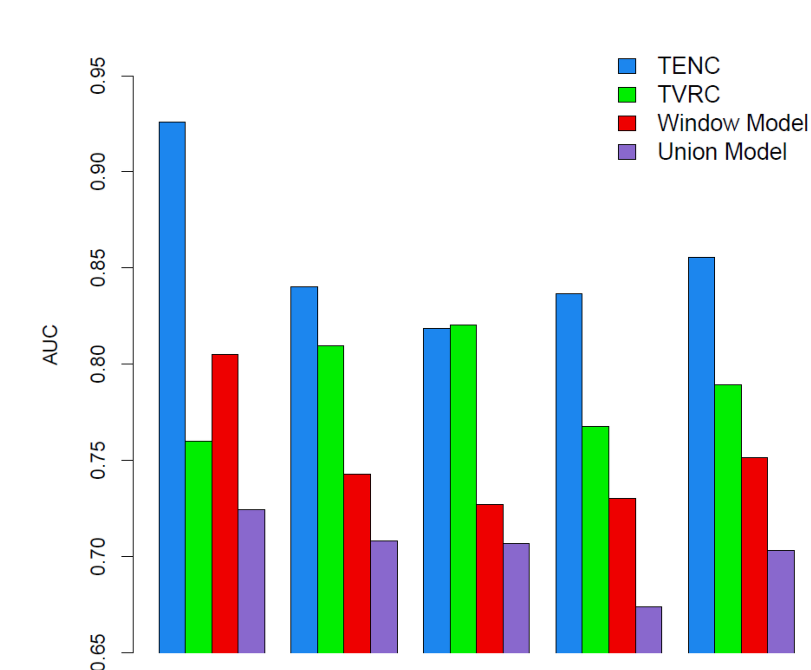

- **Phase 2: Incorporate Weights into Relational Classifier**
  - Use summary link and attribute weights in any arbitrary modified relational classifier to moderate the conditional attribute dependencies throughout the relational data graph
  - When relational attributes are considered by the model, the attribute values are weighted by the product of their attribute weight and the corresponding link weight




Weighted mode: **TVRC** → red ($\tau_3$) and **TENC** → blue($\tau_2$)

## Results: Predicting Effectiveness

- Weighting parameters θ and λ are selected using k-fold cross validation
- Models:
  - TENC: Incorporates the temporal influence of **both** links and attributes
  - TVRC: Uses temporal information on links **only**
  - Union Model: Uses unweighted summary network
  - Window Model: Uses only the immediate past
- **Main Finding:** ***TENC drastically improves model performance over all models***




## Conclusions

- Main Contributions:
  - Method to automatically annotate network with latent link and node topics for classification
  - Designed classifier to model and leverage the evolution of both **links** and **latent topics**
- Modeling the **temporal dynamics of the latent topics** results in a **significant** improvement for predicting individual effectiveness
- The results illustrate the opportunity for modeling both the time-varying communication links and the temporally evolving latent topic attributes