SCARLET: Explainable Attention based Graph Neural Network for Fake News spreader prediction

Bhavtosh Rath[1], Xavier Morales[2], and Jaideep Srivastava[1]

[1] University of Minnesota, USA
rathx082@umn.edu, srivasta@umn.edu
[2] Harvard College
xavier_morales@college.harvard.edu

PAKDD'21

211026 Chia-Chun Ho

Outline

Introduction

Related Work

Preliminaries

Proposed Approach

Experiments

Conclusion and Future work

Comments

Introduction

False information on social network

- Social network platforms like Twitter, Facebook and WhatsApp are used by millions around the world to share information and opinions.
- Often, the veracity of content shared on these platforms is not confirmed.
- This gives rise to scenarios where information having conflicting veracity, i.e. false information and its refutation, co-exist.
- Refutation can be defined as true information which fact checks claims made by a false information.

Introduction

False information spreading

- An equally important problem with fake news detection is that of preventing the impact of false information spreading.
- Techniques involve suppression of false information, as well as accelerating the spread of its refutation.
- Being able to predict the likely action of such users before they are exposed to false information is an important aspect of such a strategy.

Introduction

False information spreading

- Node identified as vulnerable to believing false information can thus
  - Be cautioned about the presence of the false information so that don't propagate it.
  - Be urged to propagate its refutation.

Introduction

False information spreading

- While optimization models based on information diffusion theories have been proposed in the past for misinformation containment.
- Recent advancements in deep learning on graphs serve as the motivation to explore false information control models.
- These models use components that exist even before false information starts spreading, namely the underlying network structure and people's historical behavioral data.

Introduction

Trust and Credibility meanings

- Trust and Credibility are important psychological and sociological concepts respectively, that have subtle differences in their meanings.
- Trust
  - represents the confidence one person has in another person.
- Credibility
  - represents generalized confidence in a person based on their perceived performance record.

Introduction

Trust and Credibility in graph representation

- Thus, in a graph representation of a social network.
- Trust
    - Property of a (directed) edge.
- Credibility
    - Property of an individual node.

Introduction

Proposed method

- Metzger et al.* showed that the interpretation of a neighbor's credibility by a node relies on its perception of the neighbor based on their trust dynamics.

- Motivated with this idea, propose a graph neural network model that integrates people's credibility and interpersonal trust features in a social network to predict whether a node is likely to spread false information or not.

Introduction

Contribution

- Propose SCARLET, a novel user-centric using graph neural network with attention mechanism to predict whether a node will most likely spread false information, its refutation or be a non-spreader.

- Demonstrate that a person's decision to spread a false information is sensitive to its perception of neighbor's credibility, and this perception is a function of trust dynamics with the neighbors.

- To best of authors' knowledge, this's the first model being evaluated on real world Twitter datasets of co-existing false and refutation information.

Related Work

of false information spreading

- Credibility perception to be an important factor for believing false information.
- Interpersonal trust also played an important role win rumor transmission.
- Many computational techniques to combat false information spreading have been explored over the past decade, as summarized by Sharma et al.
- Most models rely on generating relevant features from the information that help distinguish false information from true.

Related Work

of false information spreading

- Budak et al. proposed an optimization strategy to identify false information spreaders in a network who, when convinced by its refutation, would minimize the number of people receiving the false information.
- Nguyen et al. proposed greedy approaches to a similar problem of limiting the spread of false information in social networks.
- More recently, Tong et al. studied the problem as a multiple cascade diffusion problem.

Preliminaries

Interpersonal Trust-based features: Global Trust T_r^G

- Global trust are trust scores that are computed on the directed follower-follower network around information spreaders.
- Individual's trust score is sensitive to changes in the network structure.
- Using the Trust in Social Media (TSM) algorithm, quantify the likelihood of trusting others and being trusted by others.

Preliminaries

Interpersonal Trust-based features: Global Trust T_r^G

- TSM algorithm uses a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ as input, together with a specified convergence criteria, and computes trustingness and trustworthiness scores:
- Trustingness:

$$ti(v) = \sum_{\forall x \in out(v)} \left(\frac{w(v,x)}{1 + (tw(x))^s} \right)$$

- Trustworthiness:

$$tw(u) = \sum_{\forall x \in \mathrm{in}(u)} \left(\frac{w(x,u)}{1 + (ti(x))^s} \right)$$

- $u, v, x \in \mathcal{V}$: nodes
- $w(v,x)$: weight of edge from v to x
- $out(v)$: set of out-edges of v
- $in(u)$: set of in-edges of u
- s: involvement score of the network

Preliminaries

Interpersonal Trust-based features: Local Trust T_r^L

- Computed based on the retweeting behavior of an individual.
- It's termed local because the trust score depends on node's behavior, and not on the network structure.
- Consider the proxy for trusting others as fraction of tweets of x that are retweets (RT_x) denoted by $\sum_{\forall i \in t} \{1 \text{ if } i = RT_x \text{ else } 0\}/n(t)$.
- Consider the proxy for trusted by others as the average number of times x's tweets are retweeted ($n(RT)$) denoted by $\sum_{\forall i \in t} i_{n(RT_x)}/n(t)$.

Preliminaries

Credibility-based features: User-based Credibility C_r^U

- Extracted from user metadata of nodes in the network.
- Registration age (U1): time that has transpired since a user created their account. Older accounts tend to be associated with more credible users.
- Overall activity count (U2): Activity or statuses count is the number of tweets issued by a user. Low credibility is associated with users who have less activity on their timeline.
- Is verified (U3): This label suggests whether a user account is marked as authentic or not by Twitter. Verified accounts are more likely to be credible.

Preliminaries

Credibility-based features: Content-based Credibility C_r^C

- Obtained by aggregating a user's timeline activity.
- Do not make a distinction between information that is specifically related to news or not, as that process would require manually assessing newsworthiness of the tweets.
- Emotions conveyed by user (M1): represent positive or negative sentiments associated with a tweet. Strong sentiments are usually associated with non-credible users.
- Level of uncertainty (M2): quantified as the fraction of user's tweets that are questioning in nature. Tweet with a high level of uncertainty tend to be less credible.
- External source citation (M3): quantified as the fraction of user's tweets that cite an external URL. Tweets which cite URLs tend to be more credible.

Proposed Approach

Problem formulation

- Let $\mathscr{G}(\mathscr{V}, \mathscr{E})$ be a directed social network containing false information spreaders ($\mathscr{V}_F$), refutation information spreaders ($\mathscr{V}_T$) and non-spreaders ($\mathscr{V}_{\hat{S}_p}$) at a time instance $t(\{\mathscr{V}_F \cup \mathscr{V}_T \cup \mathscr{V}_{\hat{S}_p}\}) \subset \mathscr{V}$.
- By assigning importance score using global (T_r^G) and local (T_r^L) trust features ($T_r = T_r^G \parallel T_r^L$), and aggregating user-based (C_r^U) and content-based (C_r^C) credibility features ($C_r = C_r^U \parallel C_r^C$) of node i and its neighborhood nodes $\mathscr{N}_i^K$ sampled till depth K.
- Predict whether i is more likely to spread false information, refutation information or be non-spreader at future time $t + \Delta t$.

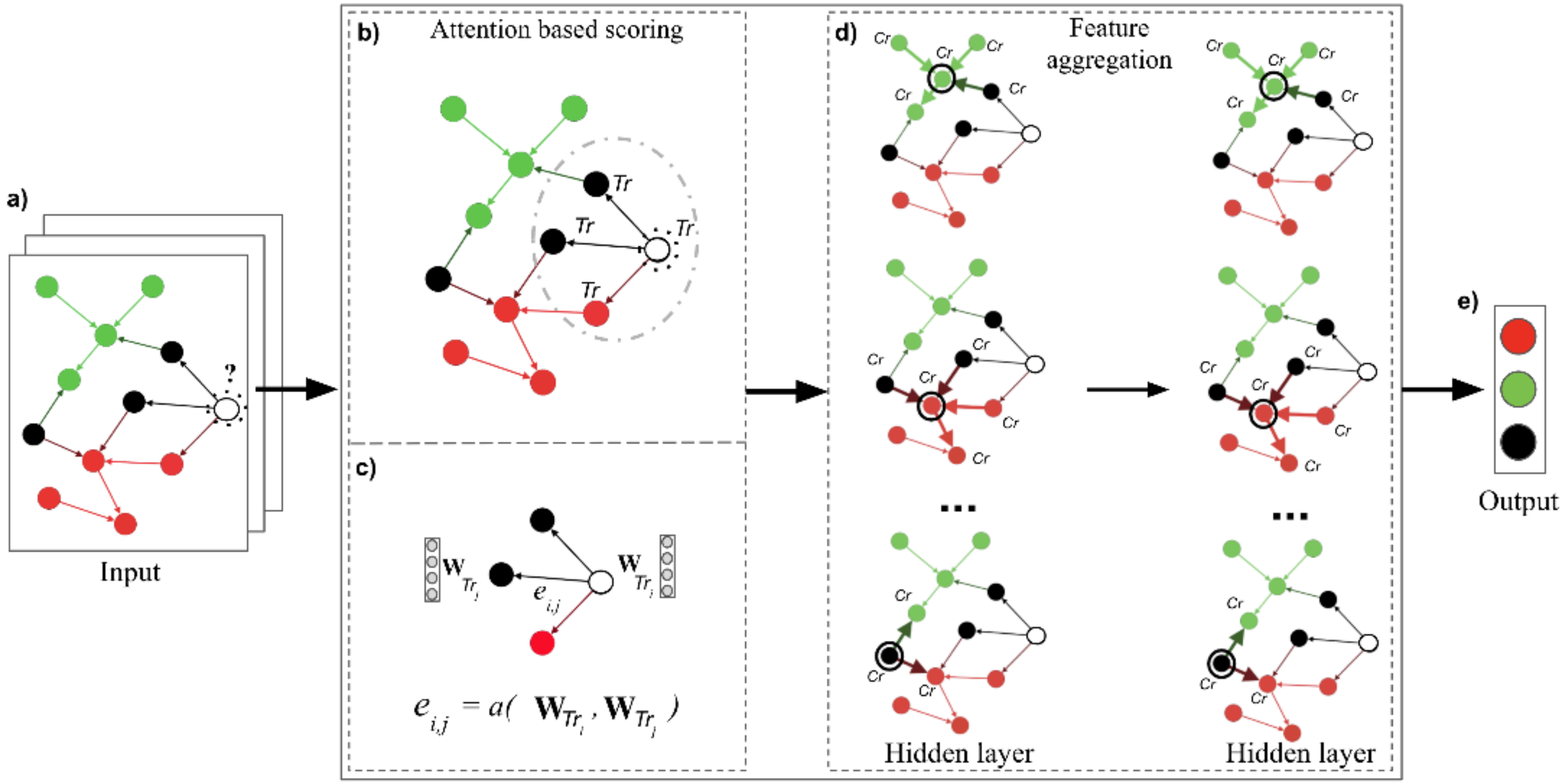
Proposed Approach

Framework Overview

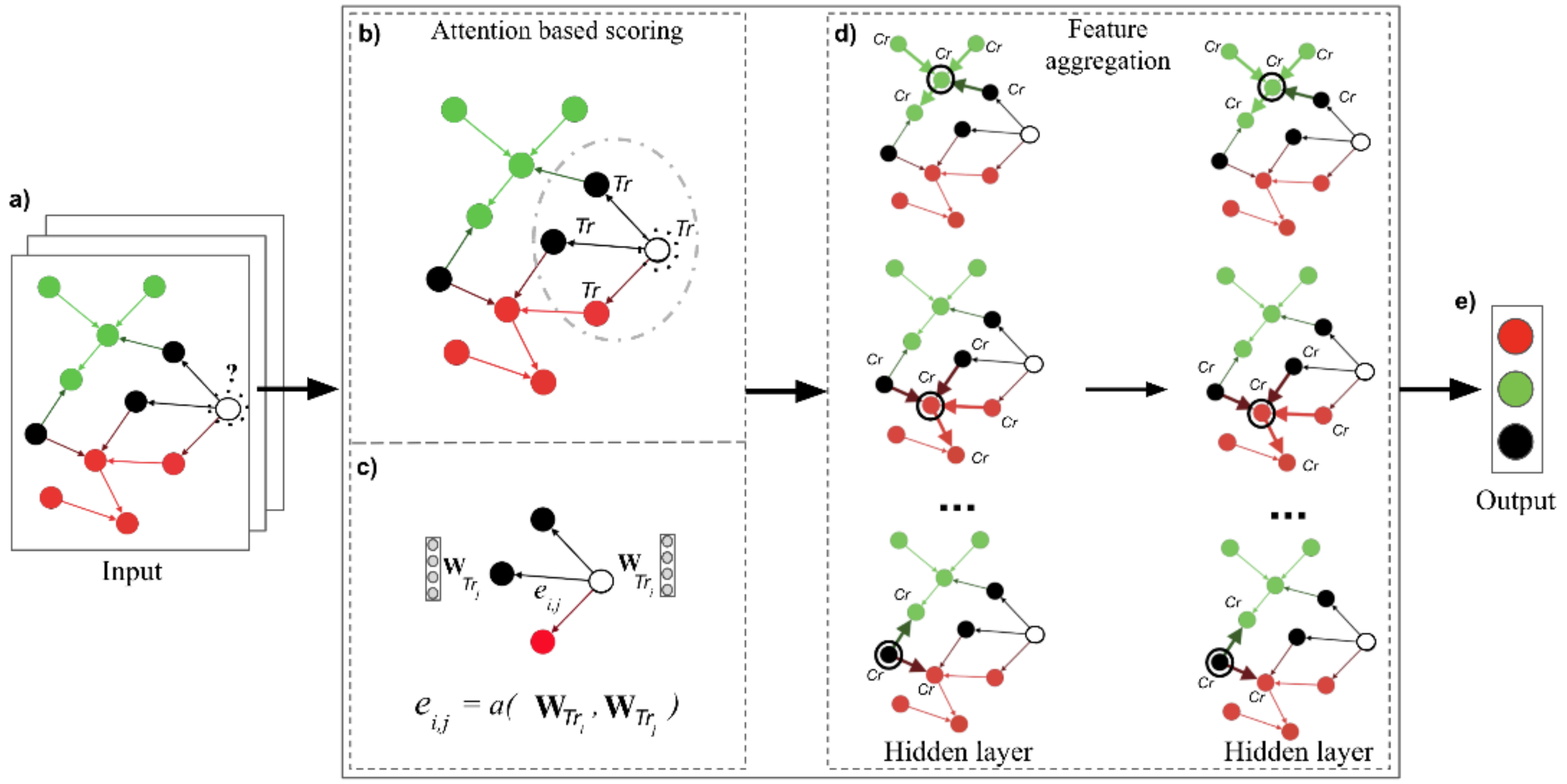
- Assign an importance score to neighborhood nodes sampled till depth K based on trust (T_r) features with attention mechanism.

- Learn representation using GCN by aggregating credibility (C_r) features proportional to importance scores assigned for the neighborhood nodes.

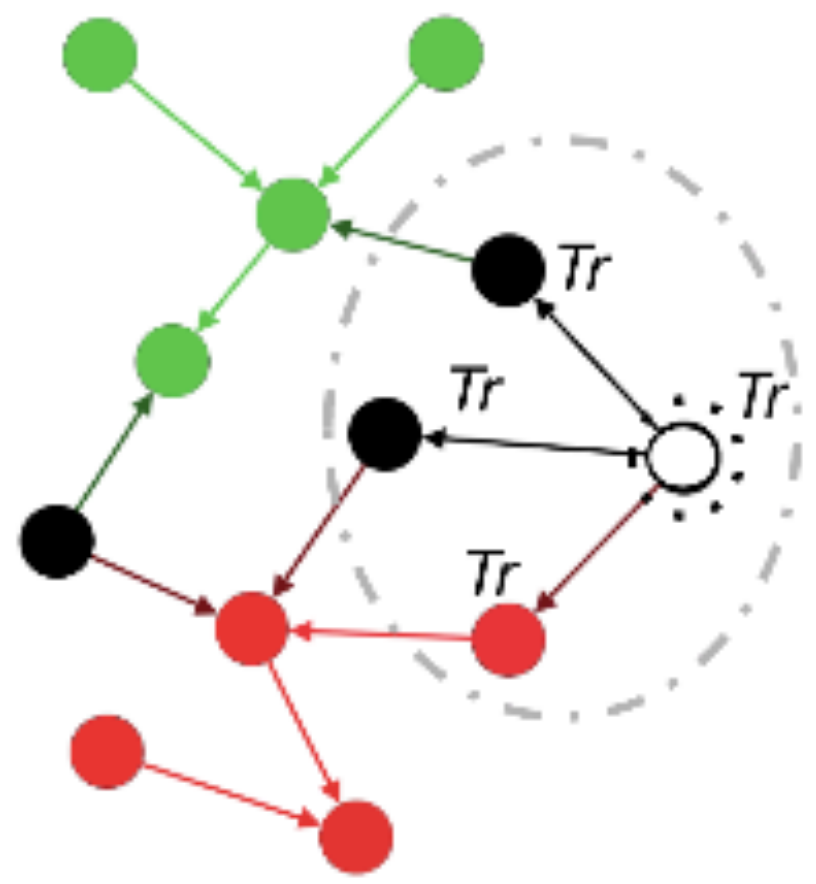
- Classification its node.



Proposed Approach
Importance score using attention

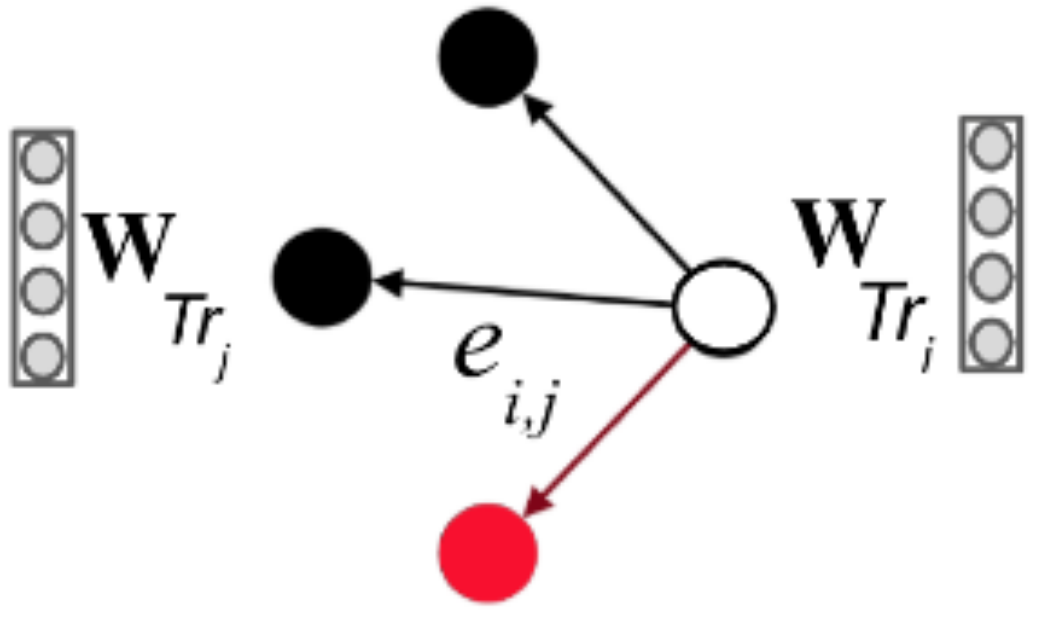


- Apply a graph attention mechanism which attends over the neighborhood of i and, based on their trust features, assigns an importance score to every j $(j \in \mathcal{N}_i)$.



- First, every node is assigned a parameterized weight matrix $(\mathbf{W})$ to perform linear transformation.

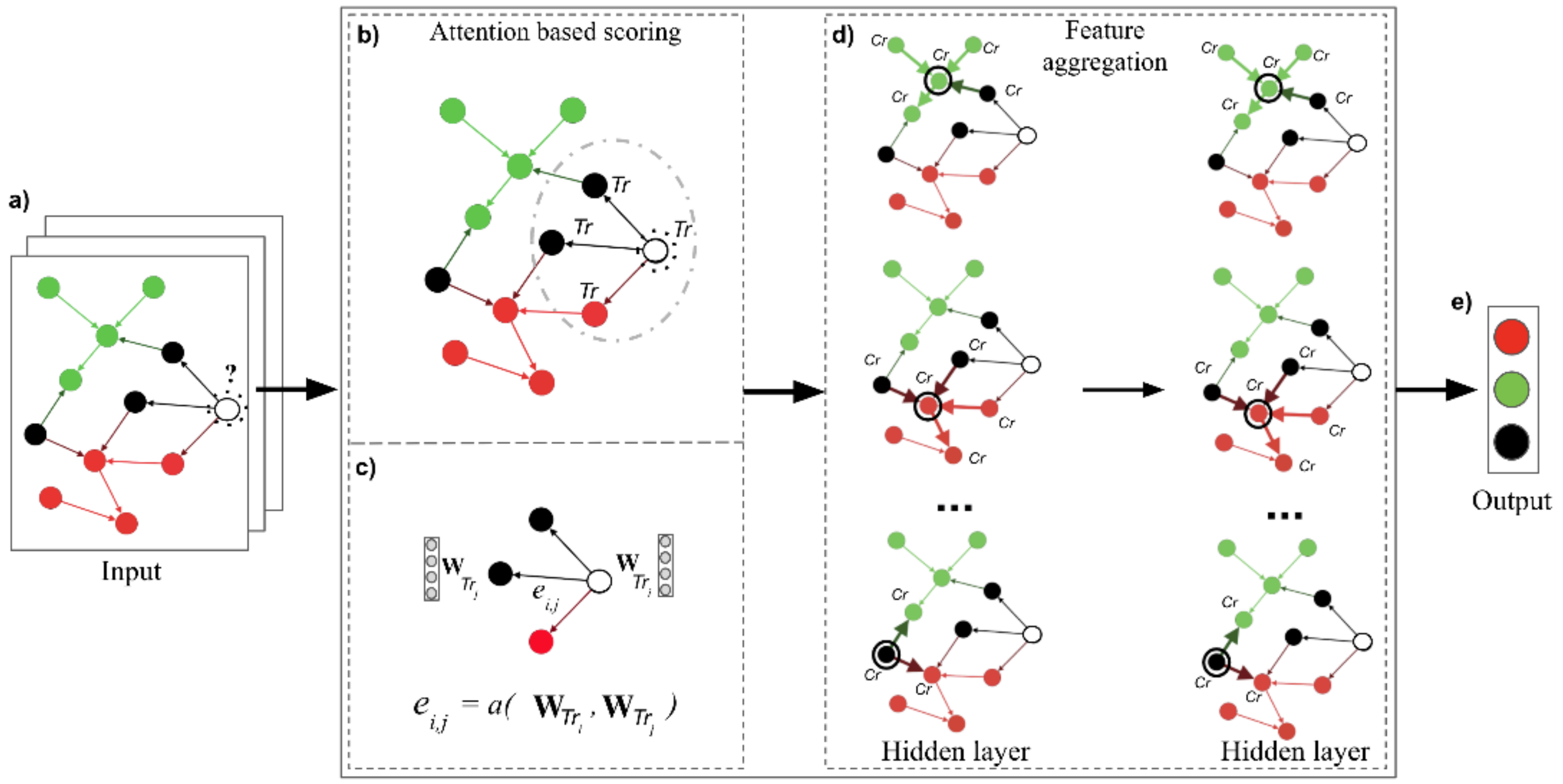
- Then self-attention is performed using a shared attention mechanism a which computes trust-based importance scores.

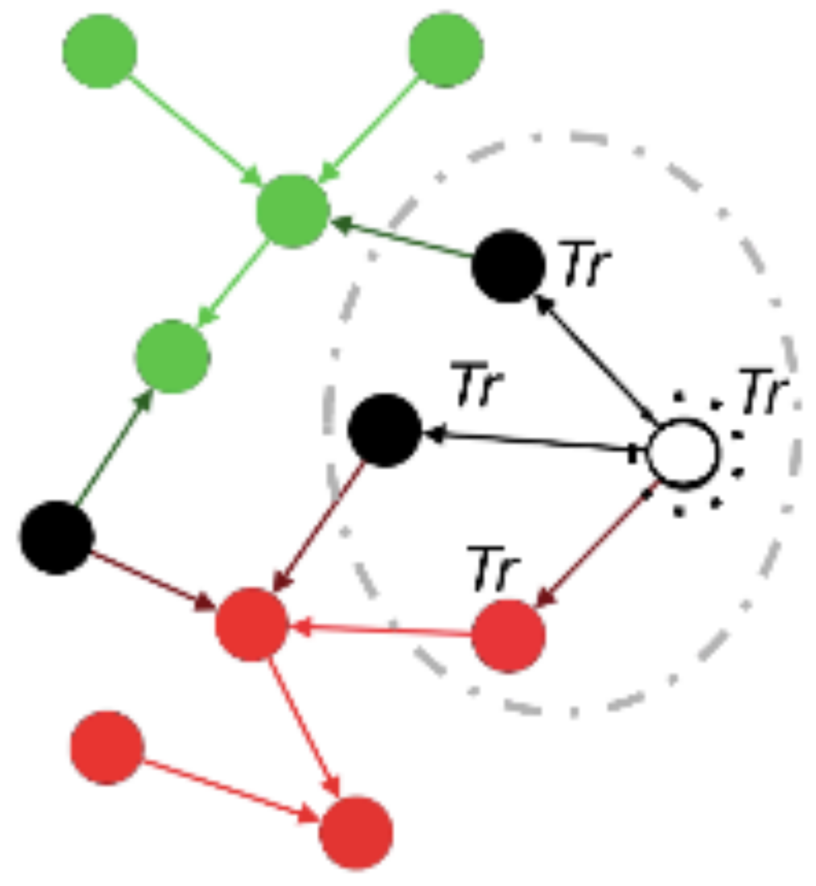


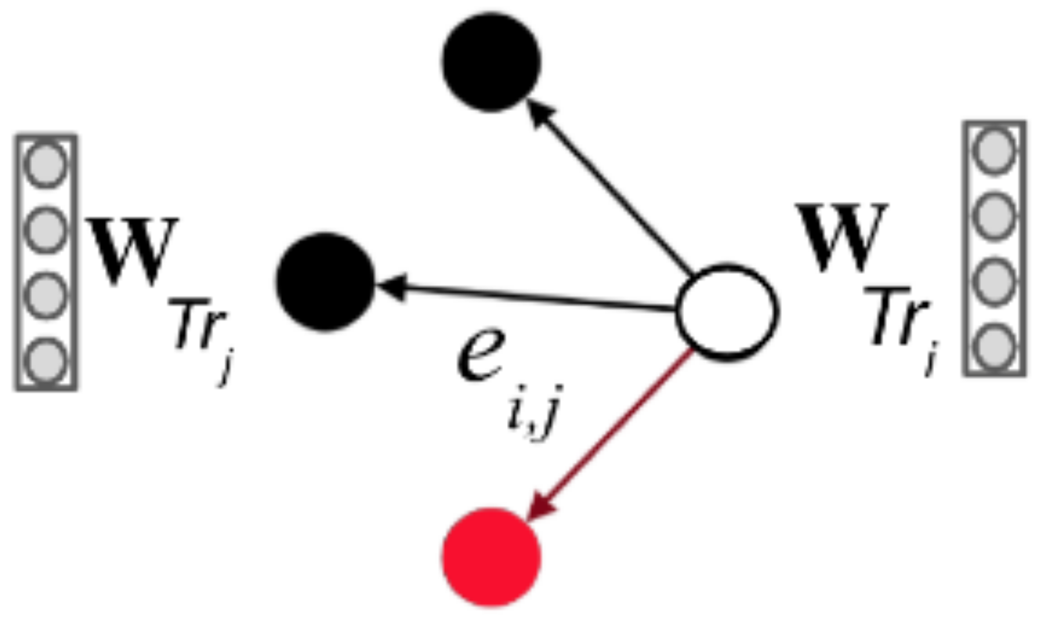
Proposed Approach

Importance score using attention

- Unnormalized trust score between i, j is represented as
  - $e_{ij} = a\left(\mathbf{W}_{Tr_i}, \mathbf{W}_{Tr_j}\right)$
  - e_{ij} quantifies j's importance to i in the context of interpersonal trust.
- Perform masked attention by only considering bodes in $\mathcal{N}_i$.
- This way aggregate features based only on neighborhood's structure.



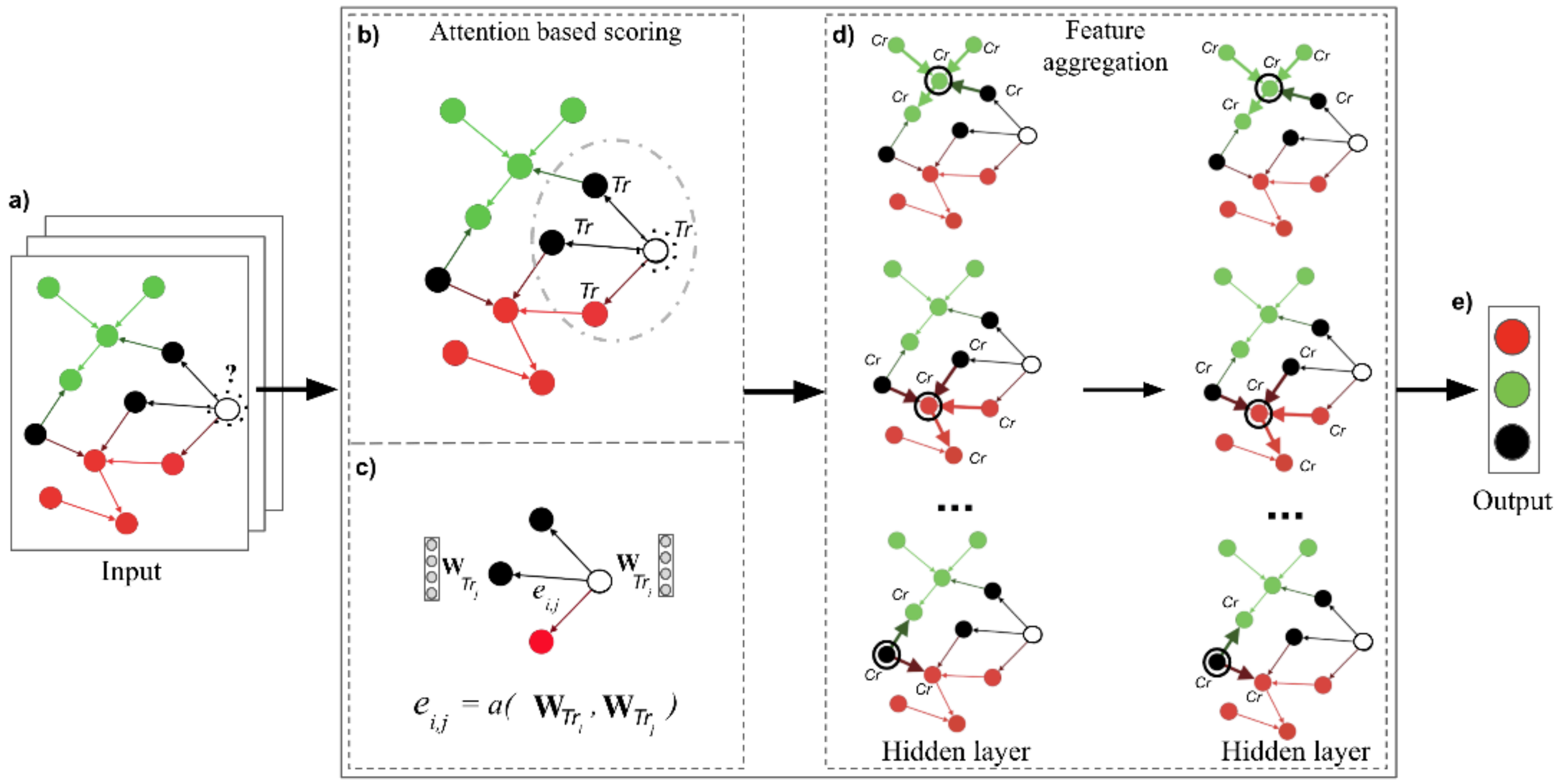


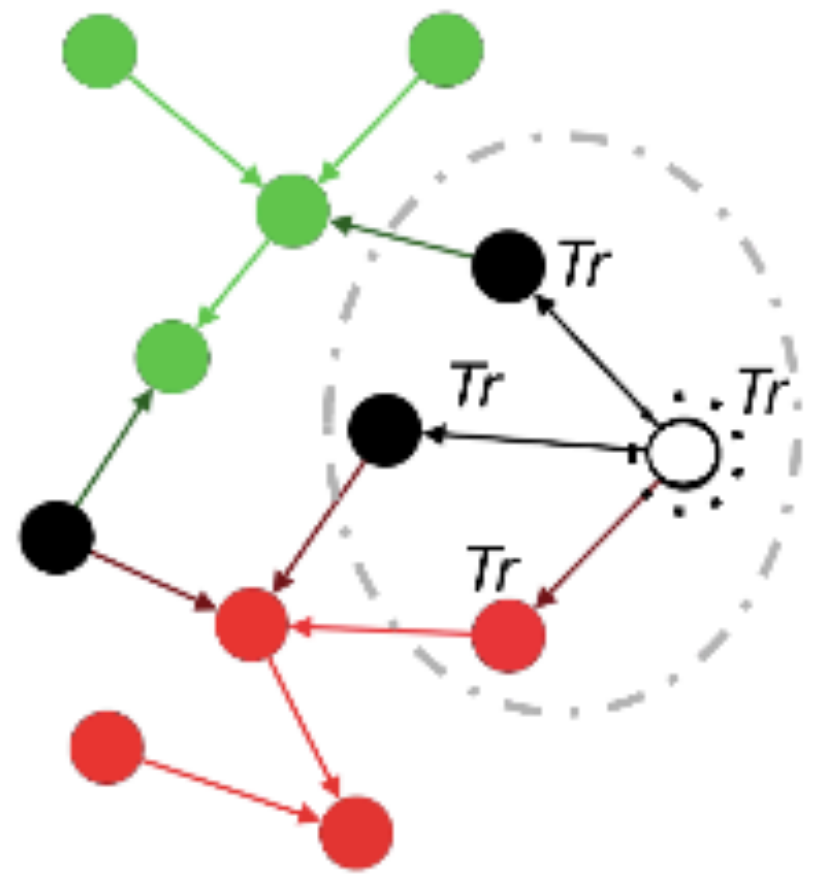


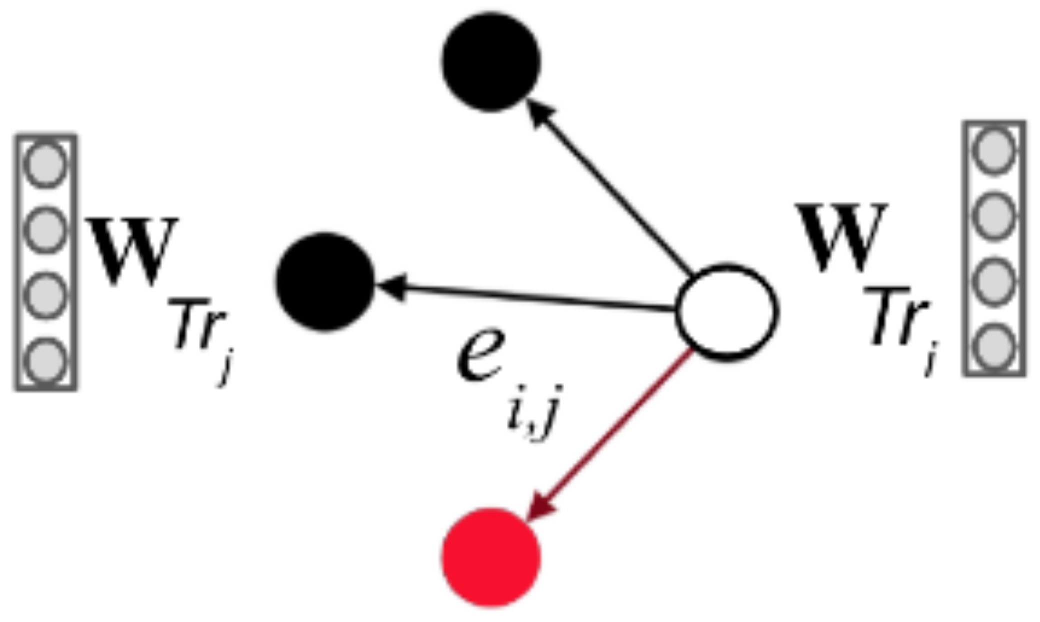
Proposed Approach

Importance score using attention

- To make the importance scores comparable across all neighbors, normalize them using the softmax function.
- $$\alpha_{ij} = \text{softmax}(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})}$$
- Attention layer a is parameterized by weight vector $\mathbf{a}$ and applied using LeakyReLU nonlinearity.
- $$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}_{Tr_i} \| \mathbf{W}_{Tr_j}]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}_{Tr_i} \| \mathbf{W}_{Tr_k}]))}$$







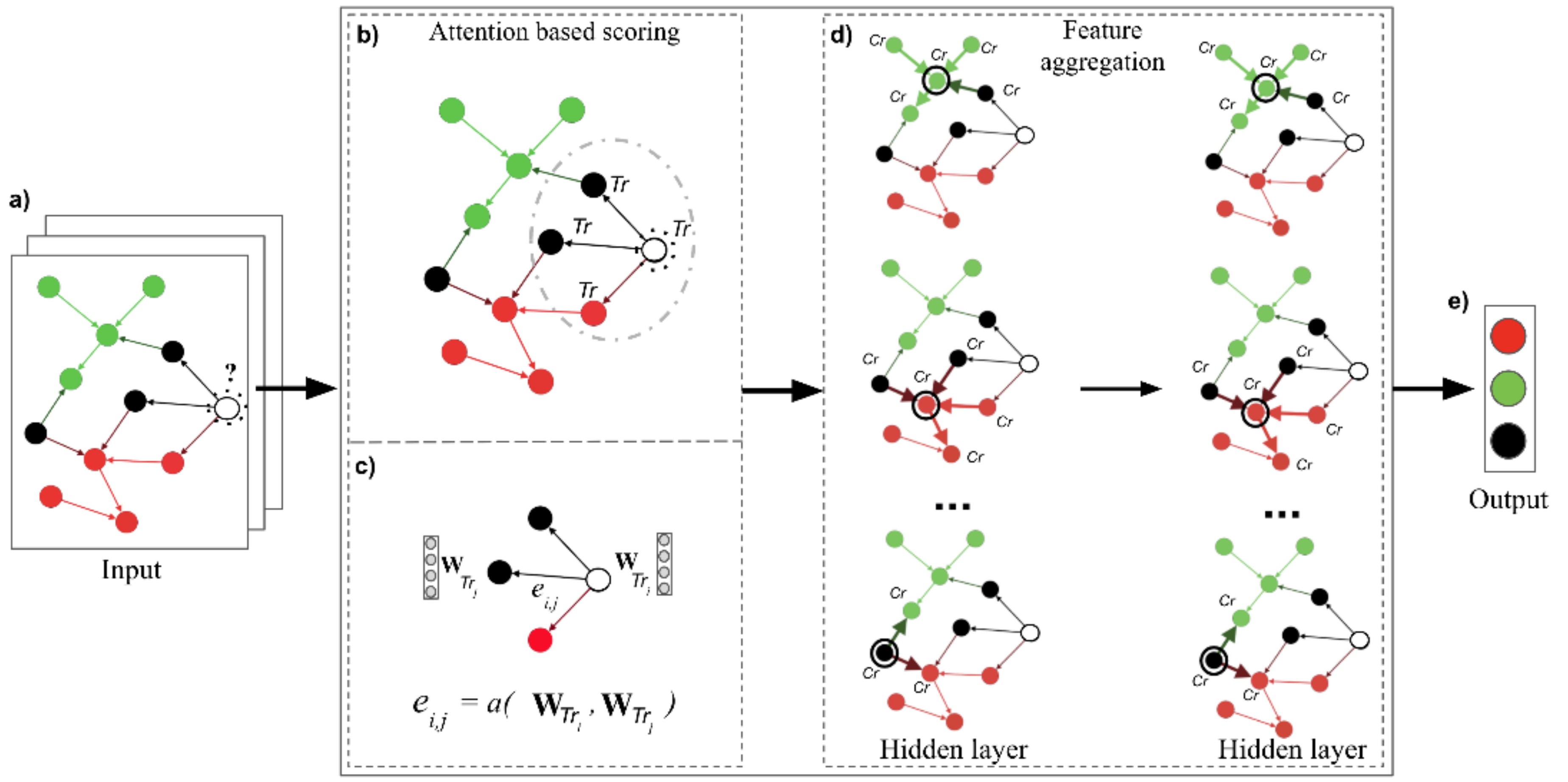
Proposed Approach

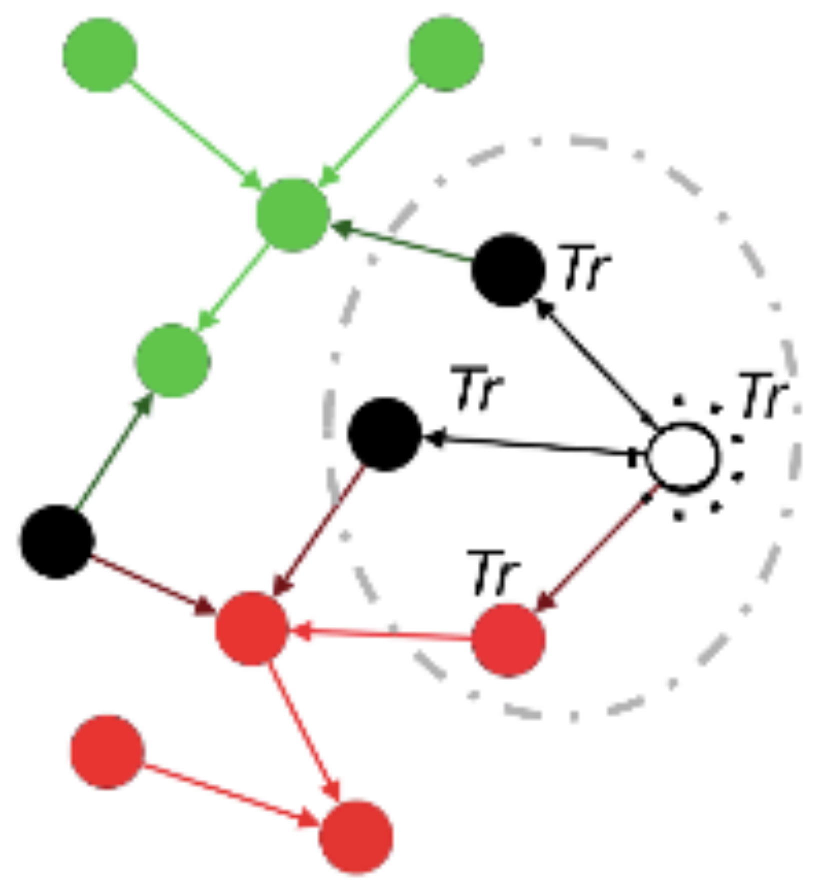
Importance score using attention

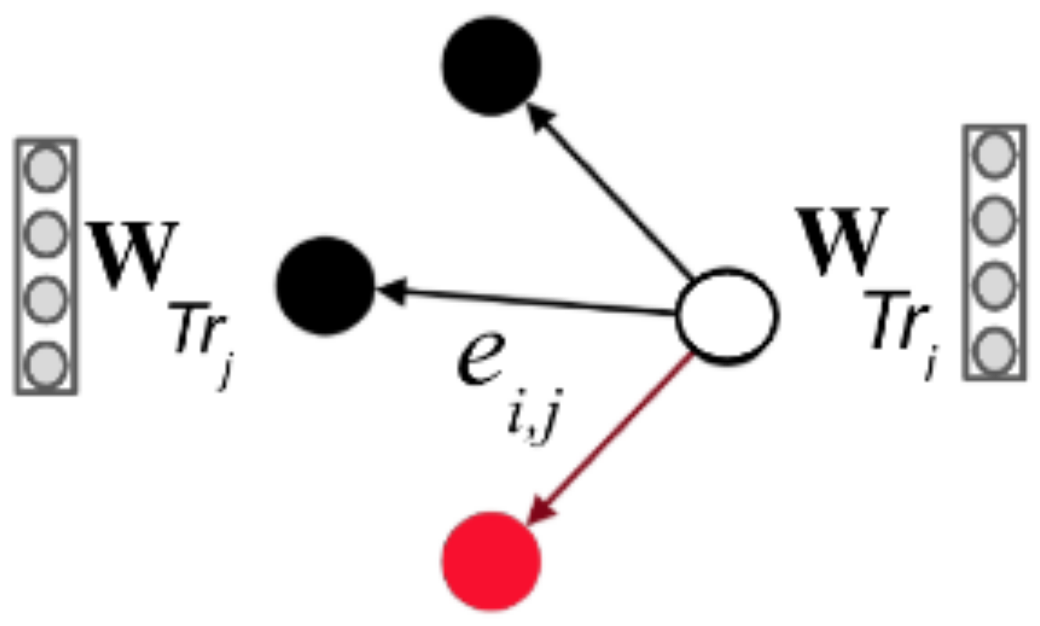
- $$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}_{Tr_i} \| \mathbf{W}_{Tr_j}]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(\mathbf{a}^T[\mathbf{W}_{Tr_i} \| \mathbf{W}_{Tr_k}]))}$$

- a_{ij} represents trust between i and j with respect to all nodes in $\mathcal{N}_i$.

- Each a_{ij} obtained for the edges is used to create an attention-based adjacency matrix $\hat{A}_{atn} = [a_{ij}]_{|\mathcal{V}| \times |\mathcal{V}|}$ which is later used to aggregate credibility features.



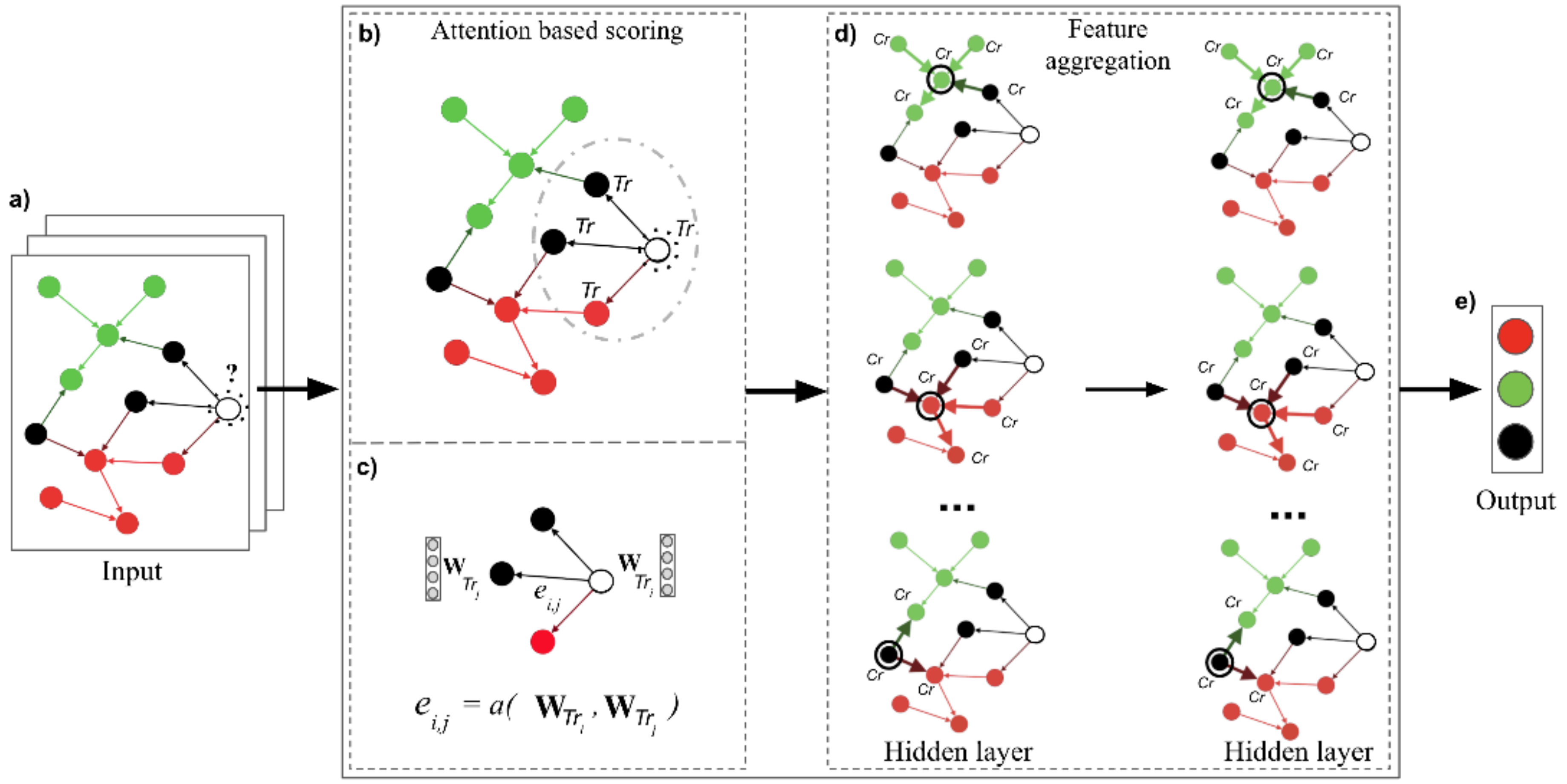


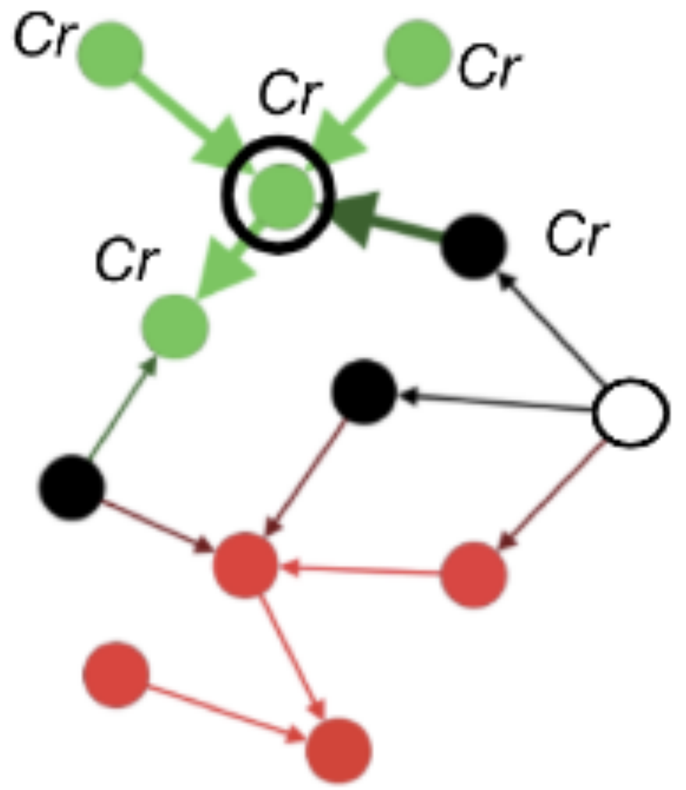


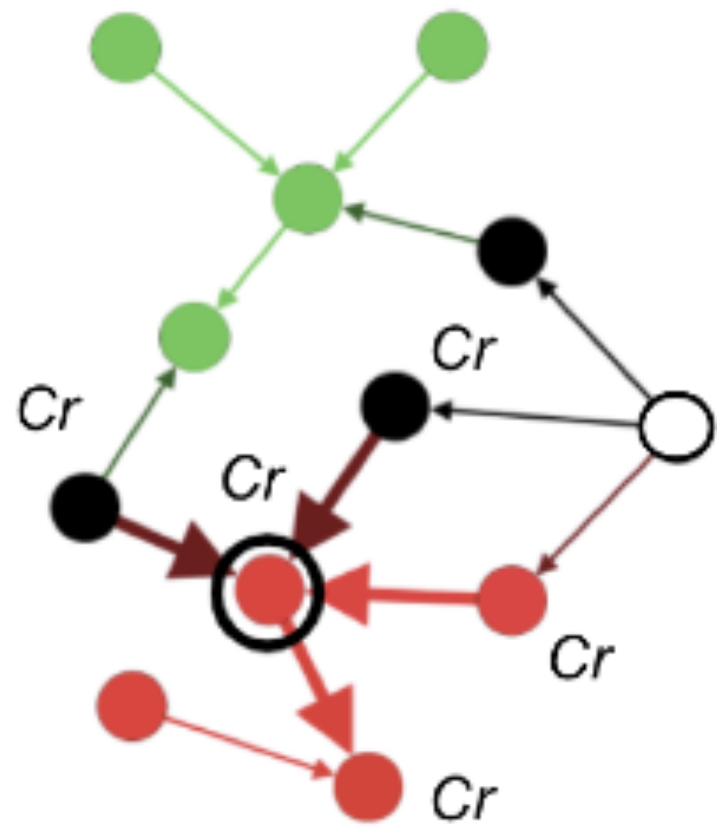
Proposed Approach

Feature aggregation

- GCN is a GNN model that efficiently aggregates features from a node's neighborhood.
- It consists of multiple NN layers where the information propagation between layers can be generalized by $H^{(l+1)} = f(H^{(l)}, A)$.
- H: hidden layer ($H^{(0)} = C_r$, $H^{(L)} = Z$)
- A: adjacency matrix representation of subgraph.
- Z: node-level output during transformation

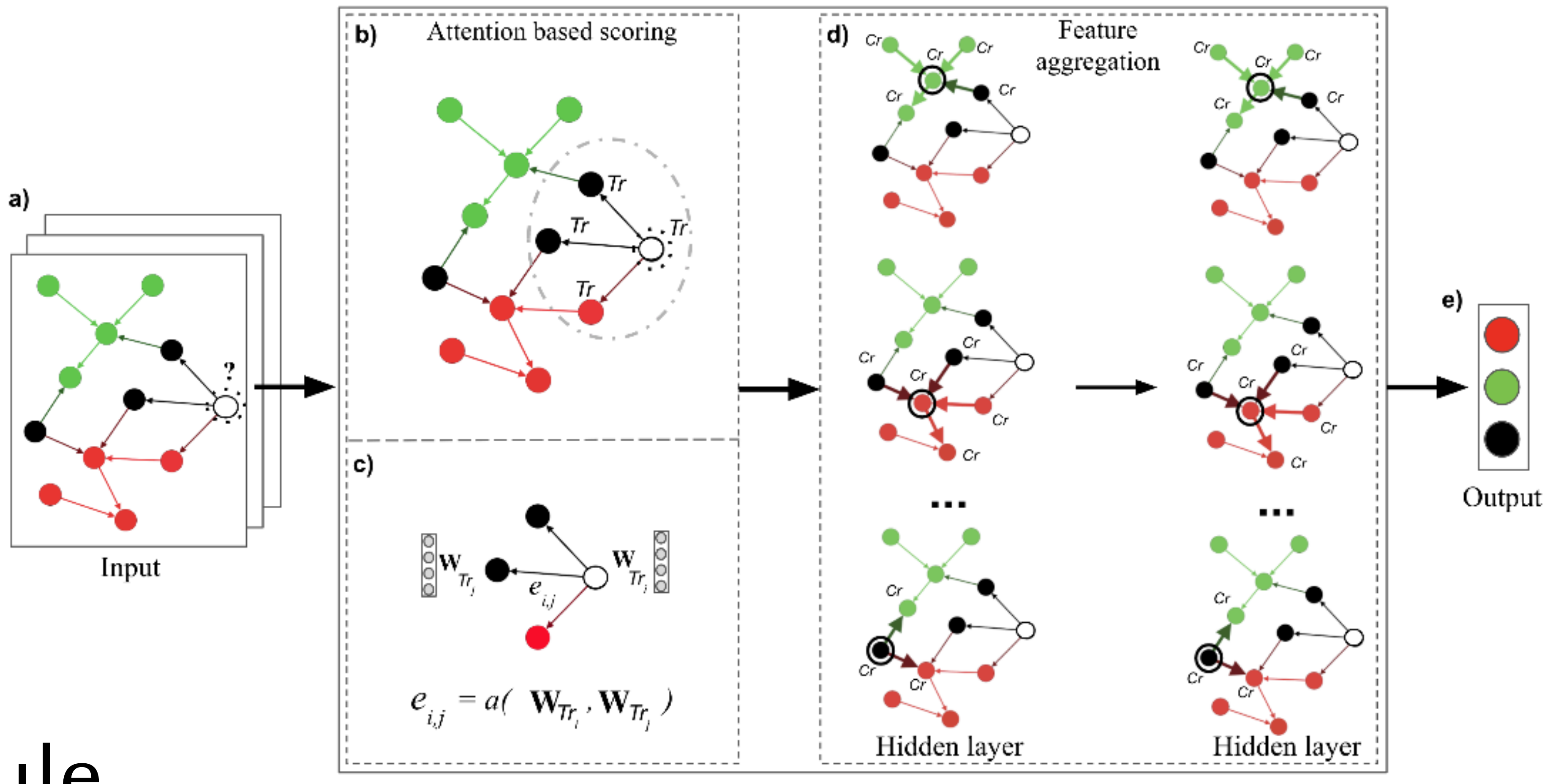




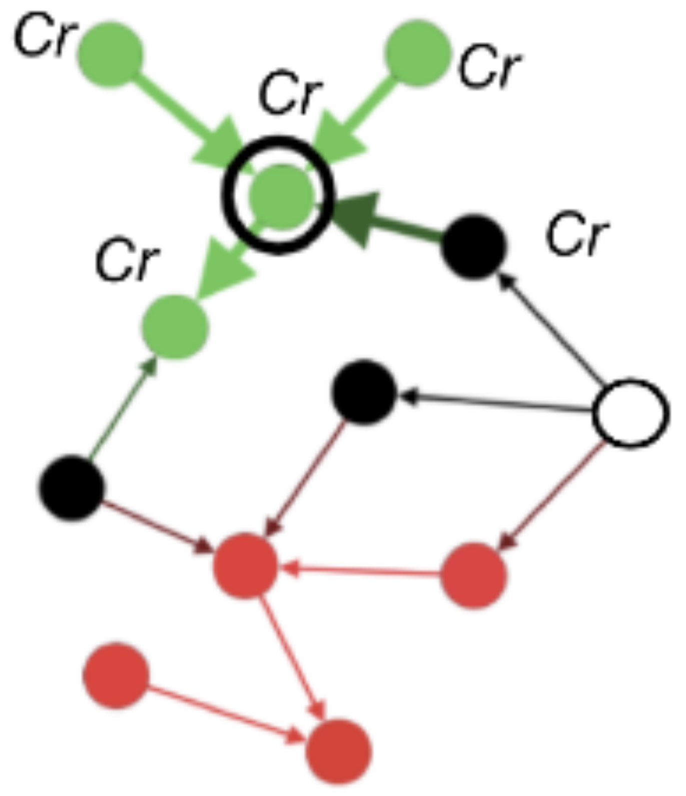


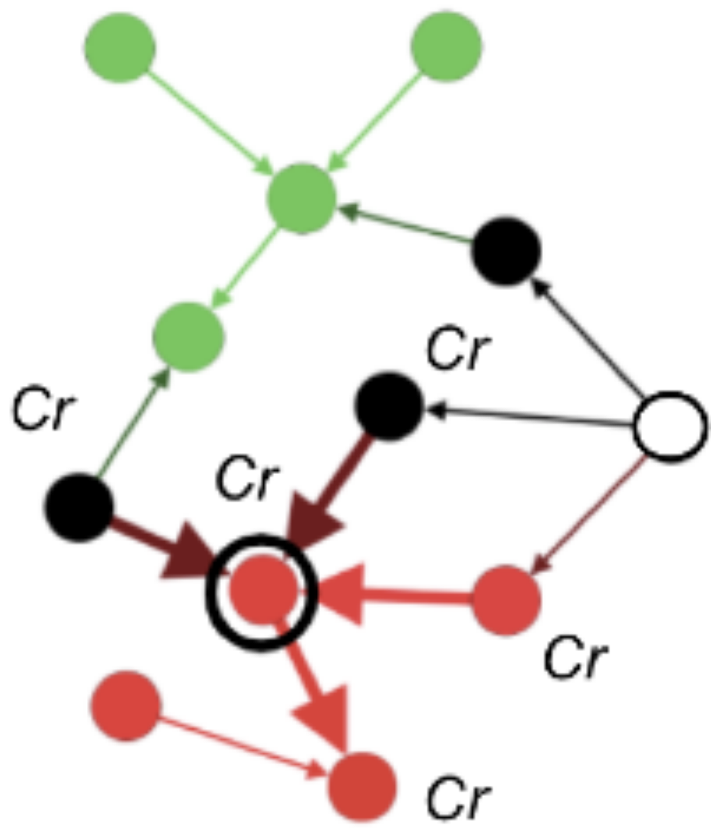
Proposed Approach

Feature aggregation



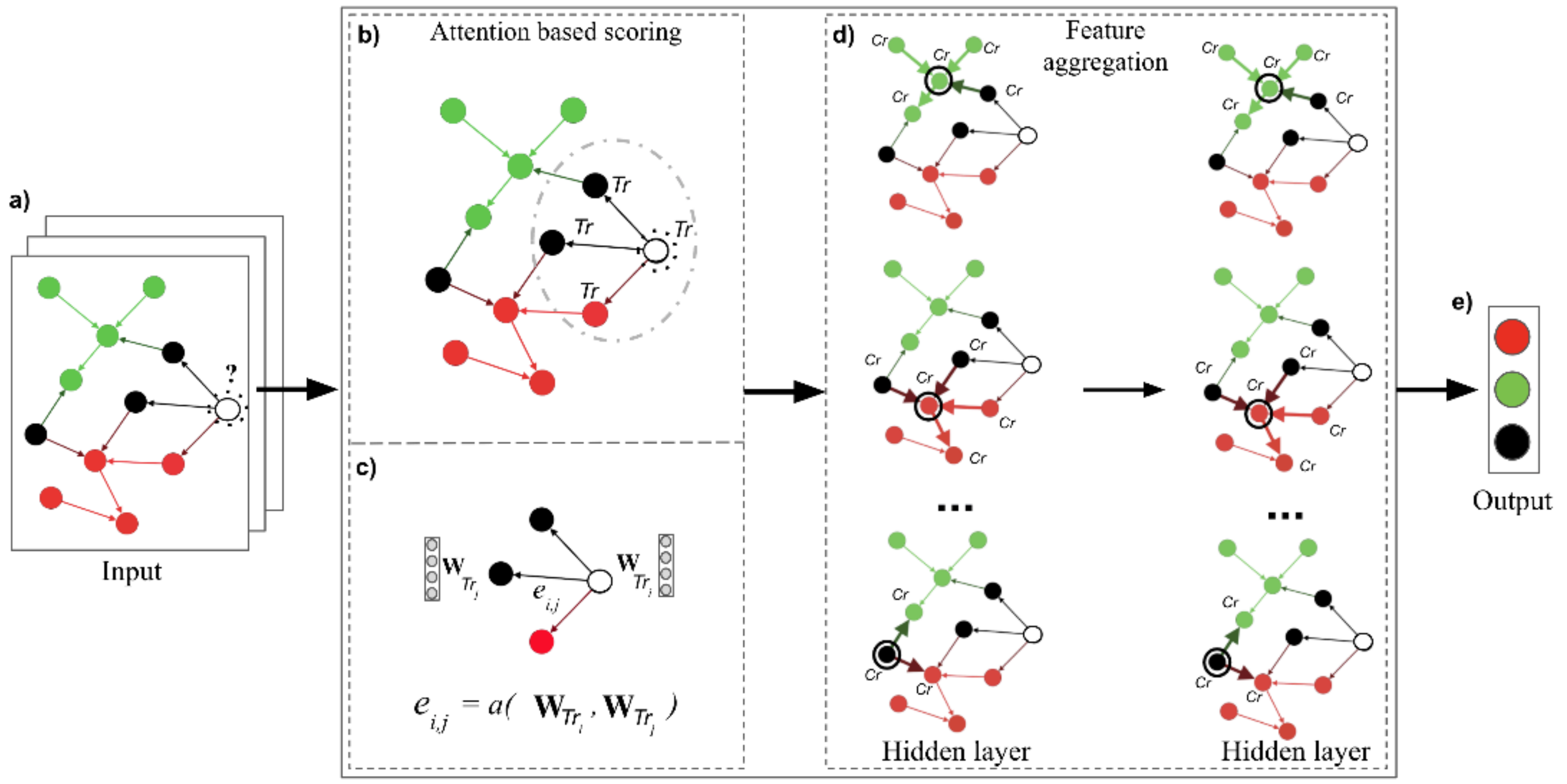
- Implement a GCN with 2 hidden layers using a propagation rule.
- $H^{(l+1)} = \sigma(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W^{(l)})$
- $\hat{A} = A + I$, ensures that include self-features during aggregation of neighbor's credibility features.
- $\hat{D}$ is the diagonal matrix of node degrees for $\hat{A}$, where $D_{ij} = \sum_j \hat{A}_{ij}$.
- Symmetric normalization of $\hat{D}$ ensures model is not sensitive to varying scale of the features being aggregated.





Proposed Approach

Node classification

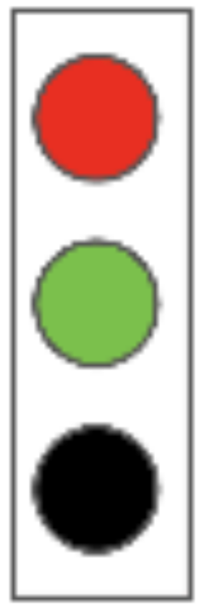


- Using credibility features and network structure for nodes in i's neighborhood, node representations are learned from the graph using a symmetric adjacency matrix with attention-based edge weights.

- Forward propagation model is applied:
  - $Z = f(X, \hat{A}_{atn}') = \mathrm{softmax}(\hat{A}\,\mathrm{ReLU}(\hat{A}XW^{(0)})W^{(1)})$

- X: credibility features

- Classification is performed using the following cross entropy loss:
  - $$\mathcal{L} = \sum_{l \in \mathcal{Y}_L} \sum_{f \in Cr} Y_{lf} \ln Z_{lf}$$



Experiments

Data collection

- Evaluate proposed model using real world Twitter datasets.
- The ground truth of false information and the refuting true information was obtained from www.altnews.in, a popular fact checking website based in India.
- The source tweet related to the information was obtained directly as a tweet embedded in the website.
- From that source tweet, used the Twitter API to determine the source tweeter and retweeters (proxy for spreaders), the follower-following network of the spreaders (proxy for social network), and user activity data (100 most recent tweets) for all nodes in the network.

Experiments

Data collection

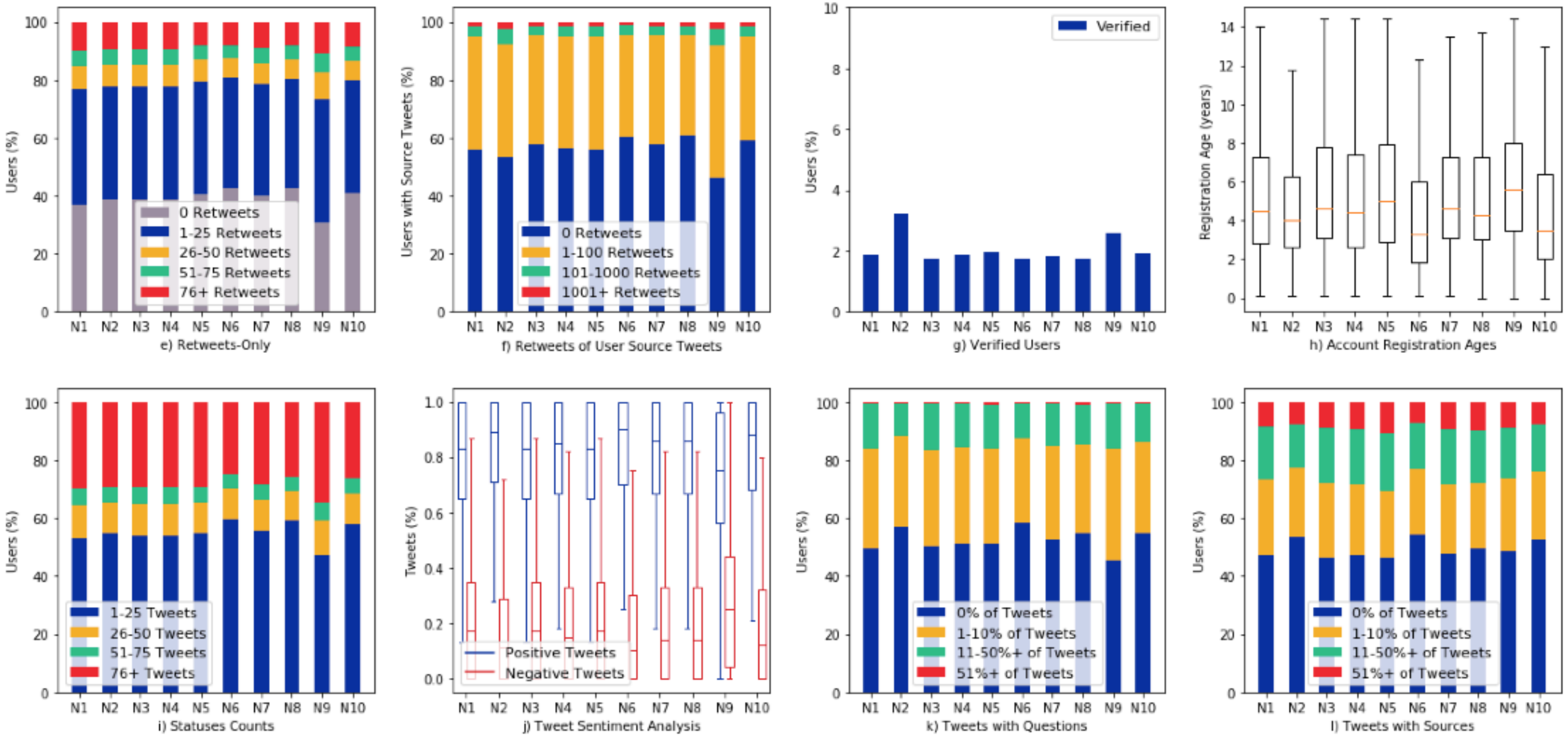
- Besides evaluating our model on the false information (F) and true information (T) spreading networks separately.
- Also evaluated proposed model on the combined information spreading networks (F ∪ T).
- Details regarding the number of nodes, edges, spreaders for the networks of 10 different news events.

	N1			N2			N3			N4			N5		
	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$
F	1,797,059	5,316,114	2,584	885,598	1,824,585	943	1,228,479	2,477,986	1,313	2,607,629	7,146,454	4,552	2,150,820	5,215,120	3,344
T	1,164,162	2,283,160	437	453,537	879,854	403	1,169,681	1,988,576	425	433,616	773,778	467	1,168,820	1,543,513	305
F ∪ T	2,677,924	7,562,503	3,017	1,230,559	2,641,513	1,337	2,198,524	4,458,228	1,738	2,900,925	7,882,019	5,015	3,019,066	6,631,032	3,627
F ∩ T	283,297	8,956	4	108,576	59,912	9	199,636	376	0	140,320	3,273	5	300,574	112,098	22

	N6			N7			N8			N9			N10		
	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert Sp\rvert$
F	2,387,610	5,356,288	3,498	627,147	1,071,120	696	2,036,162	2,876,783	894	1,197,935	2,139,912	2,317	2,174,023	4,280,962	2,323
T	1,297,371	1,727,503	481	1,166,528	2,524,907	847	1,058,482	1,513,404	489	2,999,865	6,317,032	1,833	704,006	1,314,996	741
F ∪ T	2,449,434	5,691,728	3,769	1,606,924	3,577,449	1,534	2,663,392	4,082,373	1,365	4,064,545	8,443,888	4,151	2,729,312	5,584,915	3,063
F ∩ T	1,235,547	1,379,510	212	186,751	11,131	9	431,252	305,358	20	133,255	722	1	148,717	699	1

Experiments

Data collection



Trust and credibility feature analysis from networks N1-N10

Experiments

Models and metrics: Node feature-based models

- SVM_{T_r}: applies Support Vector Machines (SVM) on node's trust based features T_r to find an optimal classification threshold.
- SVM_{C_r}: applies SVM on node's credibility based features C_r.
- SVM_{T_r,C_r}: applies SVM by combining node's trust based and credibility based features.

Experiments

Models and metrics: Network structure-based models

- LINE: applies the Large-scale Information Network Embedding as a transduction representation learning baseline.
    - Node embeddings are generated after optimization is performed on the entire graph structure.

Experiments

Models and metrics: Network structure + Node feature-based models

- SAGE_{T_r}: GraphSAGE serves as the inductive learning baseline where node embeddings are generated by aggregating T_r features from neighborhoods.
- SAGE_{C_r}: GraphSAGE to aggregating C_r features from neighborhoods.
- SAGE_{T_r,C_r}: GraphSAGE to aggregating both T_r and C_r features from neighborhoods.
- GCN_{T_r}: applies GCN to aggregating T_r features from neighborhoods.
- GCN_{C_r}: applies GCN to aggregating C_r features from neighborhoods.
- GCN_{T_r,C_r}: applies GCN to aggregating both T_r and C_r features from neighborhoods.

Experiments

Models and metrics

- SCARLET is the proposed model in this paper, which aggregates a node neighborhood's C_r features based on attention based importance scores assigned using T_r.
- For evaluation, did an 80-10-10 train-validation-test split of the dataset.
- Use 5-fold cross validation and common metric:
    - Accuracy, Precision, Recall, and F1 score.

Experiments

Performance evaluation

	F ($\mathcal{V}_F$)				T ($\mathcal{V}_T$)				F ∪ T ($\mathcal{V}_F$)			
	Accu.	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**
SVM_{Tr}	0.497	0.512	0.468	0.478	0.473	0.472	0.452	0.445	0.398	0.19	0.465	0.229
SVM_{Cr}	0.508	0.517	0.517	0.509	0.501	0.477	0.565	0.509	0.408	0.196	0.542	0.272
$SVM_{Tr,Cr}$	0.516	0.514	0.579	0.53	0.52	0.513	0.598	0.545	0.444	0.193	0.489	0.267
$LINE$	0.686	0.626	0.896	0.733	0.635	0.608	0.881	0.717	0.688	0.71	0.896	0.786
$SAGE_{Tr}$	0.734	0.762	0.691	0.722	0.680	0.698	0.719	0.705	0.752	0.743	0.859	0.793
$SAGE_{Cr}$	0.747	0.772	0.710	0.736	0.714	0.692	0.764	0.725	0.764	0.747	0.881	0.805
$SAGE_{Tr,Cr}$	0.779	0.831	0.720	0.763	**0.755**	**0.787**	0.732	0.755	0.785	0.764	0.878	0.814
GCN_{Tr}	0.784	0.726	0.947	0.821	0.718	0.675	0.916	0.767	0.753	0.783	0.930	0.845
GCN_{Cr}	0.800	0.742	0.953	0.834	0.731	0.697	0.906	0.773	0.762	0.786	0.940	0.851
$GCN_{Tr,Cr}$	0.824	0.774	0.942	0.848	0.743	0.702	0.916	0.783	0.776	**0.788**	0.954	0.861
$SCARLET$	**0.876**	**0.834**	**0.966**	**0.893**	0.734	0.674	**0.981**	**0.794**	**0.789**	0.785	**0.972**	**0.866**

- Due to class imbalance, under-sample the majority class to obtain balanced class distribution.
- Observe that structure only baseline performs better than feature only baselines.
  - Models that combine both node features and network structure show further improvement in performance.

Experiments
Performance evaluation

	$\mathbf{F}$ $(\mathcal{V}_F)$				$\mathbf{T}$ $(\mathcal{V}_T)$				$\mathbf{F} \cup \mathbf{T}$ $(\mathcal{V}_F)$			
	Accu.	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**
SVM_{Tr}	0.497	0.512	0.468	0.478	0.473	0.472	0.452	0.445	0.398	0.19	0.465	0.229
SVM_{Cr}	0.508	0.517	0.517	0.509	0.501	0.477	0.565	0.509	0.408	0.196	0.542	0.272
$SVM_{Tr,Cr}$	0.516	0.514	0.579	0.53	0.52	0.513	0.598	0.545	0.444	0.193	0.489	0.267
$LINE$	0.686	0.626	0.896	0.733	0.635	0.608	0.881	0.717	0.688	0.71	0.896	0.786
$SAGE_{Tr}$	0.734	0.762	0.691	0.722	0.680	0.698	0.719	0.705	0.752	0.743	0.859	0.793
$SAGE_{Cr}$	0.747	0.772	0.710	0.736	0.714	0.692	0.764	0.725	0.764	0.747	0.881	0.805
$SAGE_{Tr,Cr}$	0.779	0.831	0.720	0.763	**0.755**	**0.787**	0.732	0.755	0.785	0.764	0.878	0.814
GCN_{Tr}	0.784	0.726	0.947	0.821	0.718	0.675	0.916	0.767	0.753	0.783	0.930	0.845
GCN_{Cr}	0.800	0.742	0.953	0.834	0.731	0.697	0.906	0.773	0.762	0.786	0.940	0.851
$GCN_{Tr,Cr}$	0.824	0.774	0.942	0.848	0.743	0.702	0.916	0.783	0.776	**0.788**	0.954	0.861
$SCARLET$	**0.876**	**0.834**	**0.966**	**0.893**	0.734	0.674	**0.981**	**0.794**	**0.789**	0.785	**0.972**	**0.866**

- Observe that C_r features perform better than T_r features.
  - Because there are more number of C_r features than T_r features.
- Model performance increases when use T_r & C_r features together.

Experiments

Performance evaluation

	F ($\mathcal{V}_F$)				T ($\mathcal{V}_T$)				F ∪ T ($\mathcal{V}_F$)			
	Accu.	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**
SVM_{Tr}	0.497	0.512	0.468	0.478	0.473	0.472	0.452	0.445	0.398	0.19	0.465	0.229
SVM_{Cr}	0.508	0.517	0.517	0.509	0.501	0.477	0.565	0.509	0.408	0.196	0.542	0.272
$SVM_{Tr,Cr}$	0.516	0.514	0.579	0.53	0.52	0.513	0.598	0.545	0.444	0.193	0.489	0.267
$LINE$	0.686	0.626	0.896	0.733	0.635	0.608	0.881	0.717	0.688	0.71	0.896	0.786
$SAGE_{Tr}$	0.734	0.762	0.691	0.722	0.680	0.698	0.719	0.705	0.752	0.743	0.859	0.793
$SAGE_{Cr}$	0.747	0.772	0.710	0.736	0.714	0.692	0.764	0.725	0.764	0.747	0.881	0.805
$SAGE_{Tr,Cr}$	0.779	0.831	0.720	0.763	**0.755**	**0.787**	0.732	0.755	0.785	0.764	0.878	0.814
GCN_{Tr}	0.784	0.726	0.947	0.821	0.718	0.675	0.916	0.767	0.753	0.783	0.930	0.845
GCN_{Cr}	0.800	0.742	0.953	0.834	0.731	0.697	0.906	0.773	0.762	0.786	0.940	0.851
$GCN_{Tr,Cr}$	0.824	0.774	0.942	0.848	0.743	0.702	0.916	0.783	0.776	**0.788**	0.954	0.861
$SCARLET$	**0.876**	**0.834**	**0.966**	**0.893**	0.734	0.674	**0.981**	**0.794**	**0.789**	0.785	**0.972**	**0.866**

- LINE (structure only) performs better than feature only baselines by a substantial margin.
    - Suggests that network structure plays an important role in identifying false information spreaders. (Increase 32.9% (F) / 22.1% (T) / 54.9% (F ∪ T))

Experiments

Performance evaluation

	F ($\mathcal{V}_F$)				T ($\mathcal{V}_T$)				F ∪ T ($\mathcal{V}_F$)			
	Accu.	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**
SVM_{Tr}	0.497	0.512	0.468	0.478	0.473	0.472	0.452	0.445	0.398	0.19	0.465	0.229
SVM_{Cr}	0.508	0.517	0.517	0.509	0.501	0.477	0.565	0.509	0.408	0.196	0.542	0.272
$SVM_{Tr,Cr}$	0.516	0.514	0.579	0.53	0.52	0.513	0.598	0.545	0.444	0.193	0.489	0.267
$LINE$	0.686	0.626	0.896	0.733	0.635	0.608	0.881	0.717	0.688	0.71	0.896	0.786
$SAGE_{Tr}$	0.734	0.762	0.691	0.722	0.680	0.698	0.719	0.705	0.752	0.743	0.859	0.793
$SAGE_{Cr}$	0.747	0.772	0.710	0.736	0.714	0.692	0.764	0.725	0.764	0.747	0.881	0.805
$SAGE_{Tr,Cr}$	0.779	0.831	0.720	0.763	**0.755**	**0.787**	0.732	0.755	0.785	0.764	0.878	0.814
GCN_{Tr}	0.784	0.726	0.947	0.821	0.718	0.675	0.916	0.767	0.753	0.783	0.930	0.845
GCN_{Cr}	0.800	0.742	0.953	0.834	0.731	0.697	0.906	0.773	0.762	0.786	0.940	0.851
$GCN_{Tr,Cr}$	0.824	0.774	0.942	0.848	0.743	0.702	0.916	0.783	0.776	**0.788**	0.954	0.861
$SCARLET$	**0.876**	**0.834**	**0.966**	**0.893**	0.734	0.674	**0.981**	**0.794**	**0.789**	0.785	**0.972**	**0.866**

- GNN baselines that combine both network structure and node features show a significant improvement in performance.

- GCN models perform better than GraphSAGE models on all metric for F network, while that's not the case for T and F ∪ T networks.

  - This's because T_r & C_r features for neighborhood of refutation information spreaders and non-spreaders don't differ much from each other.

Experiments
Performance evaluation

	F ($\mathcal{V}_F$)				T ($\mathcal{V}_T$)				F ∪ T ($\mathcal{V}_F$)			
	Accu.	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**
SVM_{Tr}	0.497	0.512	0.468	0.478	0.473	0.472	0.452	0.445	0.398	0.19	0.465	0.229
SVM_{Cr}	0.508	0.517	0.517	0.509	0.501	0.477	0.565	0.509	0.408	0.196	0.542	0.272
$SVM_{Tr,Cr}$	0.516	0.514	0.579	0.53	0.52	0.513	0.598	0.545	0.444	0.193	0.489	0.267
$LINE$	0.686	0.626	0.896	0.733	0.635	0.608	0.881	0.717	0.688	0.71	0.896	0.786
$SAGE_{Tr}$	0.734	0.762	0.691	0.722	0.680	0.698	0.719	0.705	0.752	0.743	0.859	0.793
$SAGE_{Cr}$	0.747	0.772	0.710	0.736	0.714	0.692	0.764	0.725	0.764	0.747	0.881	0.805
$SAGE_{Tr,Cr}$	0.779	0.831	0.720	0.763	**0.755**	**0.787**	0.732	0.755	0.785	0.764	0.878	0.814
GCN_{Tr}	0.784	0.726	0.947	0.821	0.718	0.675	0.916	0.767	0.753	0.783	0.930	0.845
GCN_{Cr}	0.800	0.742	0.953	0.834	0.731	0.697	0.906	0.773	0.762	0.786	0.940	0.851
$GCN_{Tr,Cr}$	0.824	0.774	0.942	0.848	0.743	0.702	0.916	0.783	0.776	**0.788**	0.954	0.861
$SCARLET$	**0.876**	**0.834**	**0.966**	**0.893**	0.734	0.674	**0.981**	**0.794**	**0.789**	0.785	**0.972**	**0.866**

- SCARLET shows an increase in performance for all three networks.
- SAGE_{T_r,C_r} shows better accuracy and precision on T networks, because the specific news events on which it performed better involved religious tones, and so decision to refute them is more sensitive to neighborhood's C_r than T_r.
- Precision on F ∪ T networks is highest for GCN_{T_r,C_r}, though it is still comparable to the proposed model's performance.

Experiments

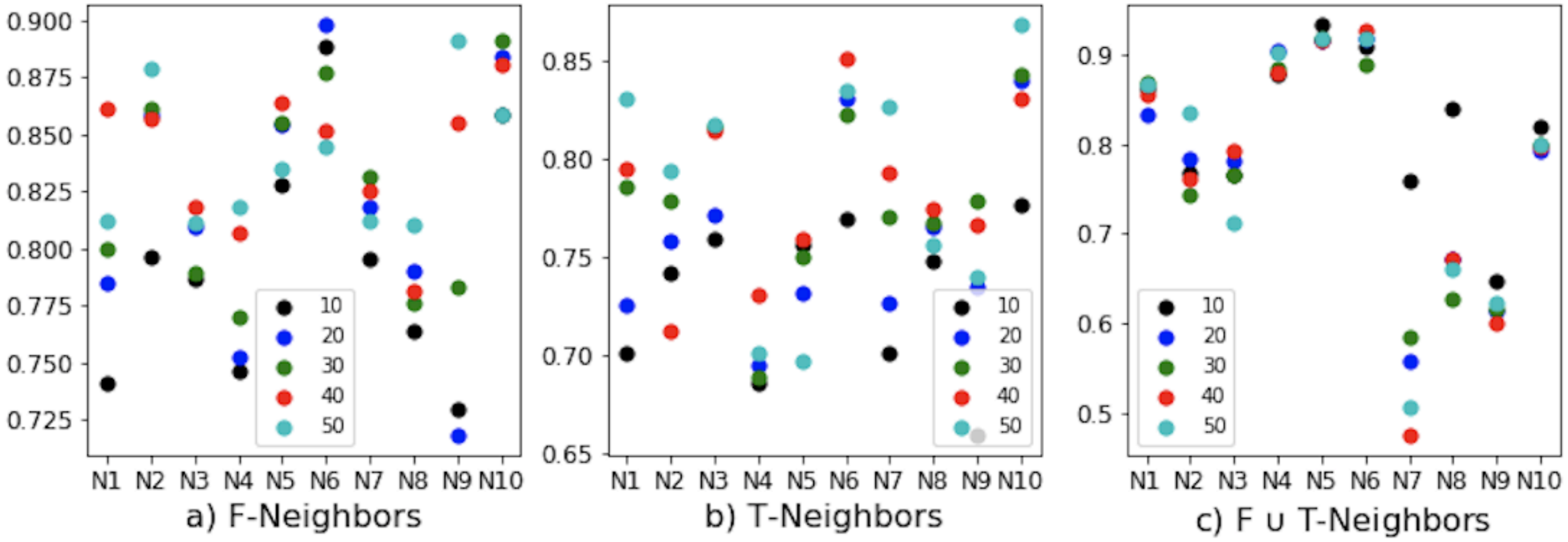
Performance evaluation

	F ($\mathcal{V}_F$)				T ($\mathcal{V}_T$)				F ∪ T ($\mathcal{V}_F$)			
	Accu.	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**	**Accu.**	**Prec.**	**Rec.**	**F1**
SVM_{Tr}	0.497	0.512	0.468	0.478	0.473	0.472	0.452	0.445	0.398	0.19	0.465	0.229
SVM_{Cr}	0.508	0.517	0.517	0.509	0.501	0.477	0.565	0.509	0.408	0.196	0.542	0.272
$SVM_{Tr,Cr}$	0.516	0.514	0.579	0.53	0.52	0.513	0.598	0.545	0.444	0.193	0.489	0.267
$LINE$	0.686	0.626	0.896	0.733	0.635	0.608	0.881	0.717	0.688	0.71	0.896	0.786
$SAGE_{Tr}$	0.734	0.762	0.691	0.722	0.680	0.698	0.719	0.705	0.752	0.743	0.859	0.793
$SAGE_{Cr}$	0.747	0.772	0.710	0.736	0.714	0.692	0.764	0.725	0.764	0.747	0.881	0.805
$SAGE_{Tr,Cr}$	0.779	0.831	0.720	0.763	**0.755**	**0.787**	0.732	0.755	0.785	0.764	0.878	0.814
GCN_{Tr}	0.784	0.726	0.947	0.821	0.718	0.675	0.916	0.767	0.753	0.783	0.930	0.845
GCN_{Cr}	0.800	0.742	0.953	0.834	0.731	0.697	0.906	0.773	0.762	0.786	0.940	0.851
$GCN_{Tr,Cr}$	0.824	0.774	0.942	0.848	0.743	0.702	0.916	0.783	0.776	**0.788**	0.954	0.861
$SCARLET$	**0.876**	**0.834**	**0.966**	**0.893**	0.734	0.674	**0.981**	**0.794**	**0.789**	0.785	**0.972**	**0.866**

- More importantly, SCARLET in the F ∪ T network observe highest accuracy and F1 scores of 78.9% and 86.6%.
  - Thus supporting proposed hypothesis that false information spreading is very sensitive to trust and credibility.

Experiments

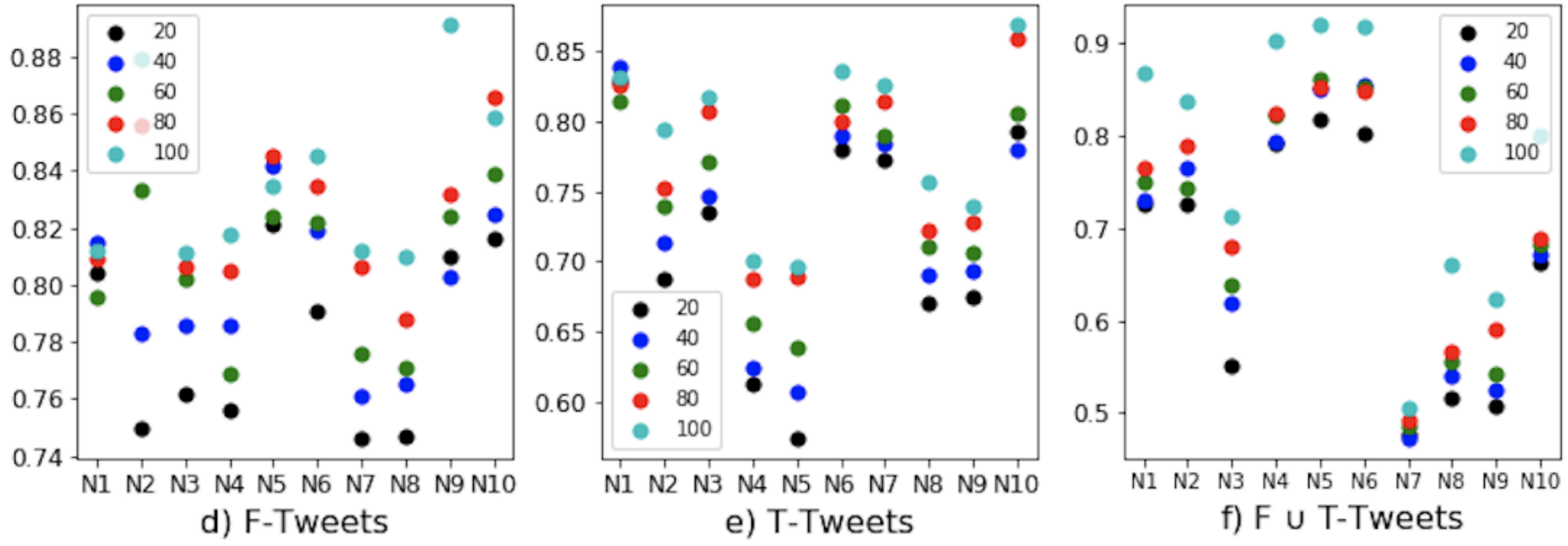
Sensitivity analysis: Neighbors



- Evaluated proposed model on n-neighbors, where n = 10, 20, 30, 40, 50.
- Observe that model performance is not very sensitive to varying neighborhood size.
  - Have only the immediate follower-following network (sampling depth=1).
  - Unable to entirely capture meaningful dynamics (i.e. the decision to retweet might depend less on the immediate neighbors, and more on the source tweeter).

Experiments

Sensitivity analysis: Neighbors



- Evaluated proposed model on the n-most recent timeline tweets, where n = 20, 40, 60, 80.
- Observe that for all three networks, prediction performance tends to increase as the number of timeline tweets used to aggregate features increases.
    - Using more behavioral data helps model to estimate trust and credibility features better.

Conclusion and Future work

- Proposed SCARLET, an attention-based explainable GNN model to predict whether a node is likely to spread false information or not.
- Model learns node embeddings by first assigning trust-based importance scores and then aggregating its neighborhood's credibility features proportionally.
  - Makes this model different from most existing research is that it doesn't rely on features extracted from the information itself.
  - Thus it can be used to predict spreaders even before information spreading begins.
- Would like to analyze model on more news events comprising larger networks in order to sample and aggregate features at greater sampling depths.

Comments
of SCARLET

- Propose concept with trust and credibility in social network.
- Using attention mechanism to compute importance score that aggregate neighborhood features proportionally.
- Without content-based information.
- In experiment, unclearly to explain F, T, F ∪ T network.