

Leveraging Joint Interactions for Credibility Analysis in News Communities

Subhabrata Mukherjee and Gerhard Weikum
Max Planck Institute for Informatics

CIKM 2015

Motivation

- Media plays a crucial role in public dissemination of information
- However, people believe there is substantial media bias in news in view of inter-dependencies and cross-ownerships of media companies and other industries (like energy)
- 4 out of 5 Americans among younger generations do not trust major news networks [Gallup poll, 2013]
- This work: [Credibility Analysis](#) of News Communities

News Community

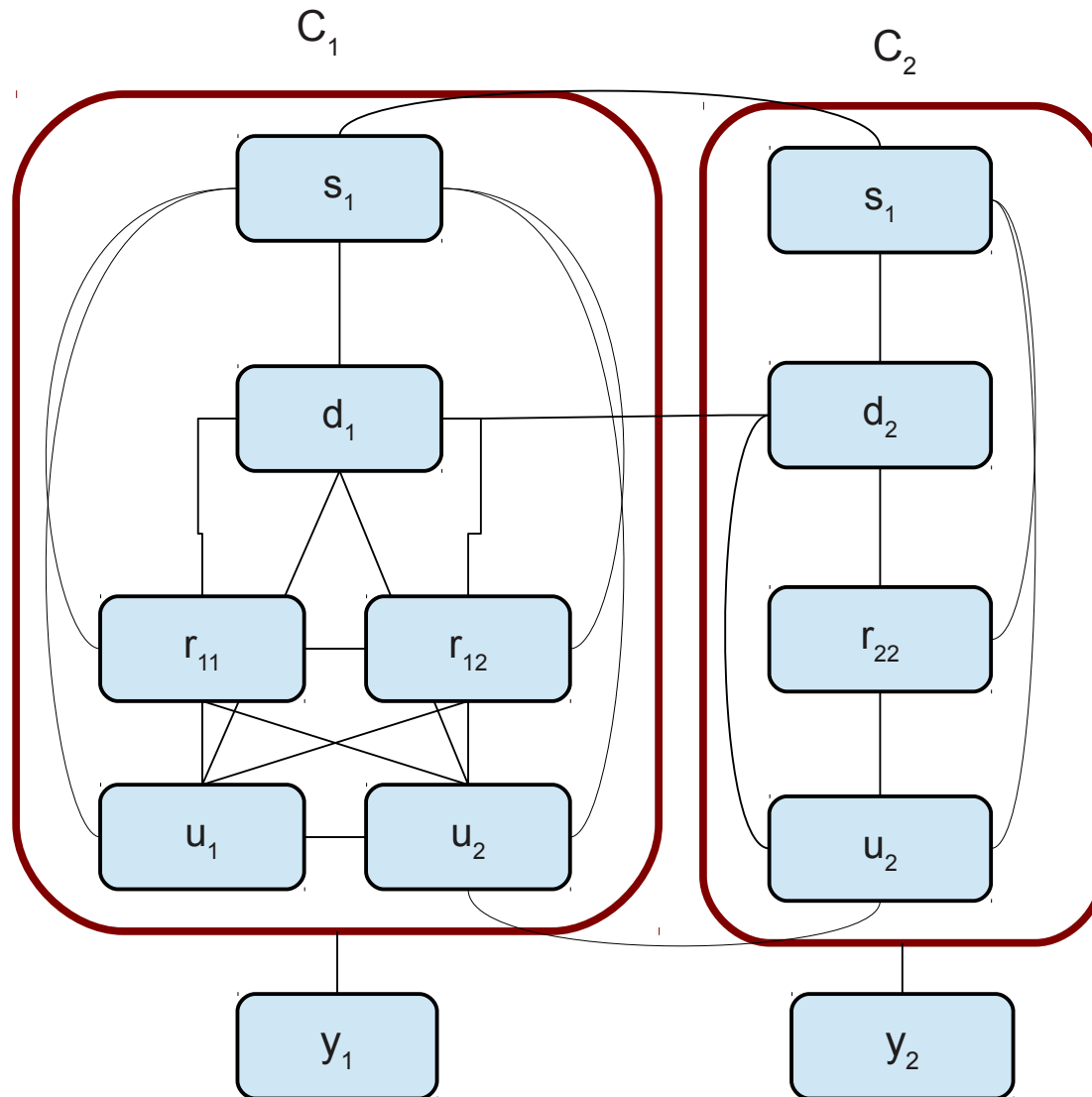
- A news community is a news aggregator site (e.g., reddit.com, digg.com, newstrust.net) where:
 - Users can give explicit feedback (e.g., rate, review, share) on the quality of news
 - Interact (e.g., comment, vote) with each other
- However, this adds user subjectivity as users incorporate their own bias and perspectives in the framework
- Controversial topics create polarization among users which influence their evaluation

Contributions

- A model to capture joint interaction between *language*, *topics*, *users* and *sources* leading to better prediction than the ones in isolation
- User *expertise*, source *trustworthiness*, language *objectivity*, topical *perspective* and article *credibility* mutually reinforce each other
- A supervised Conditional Random Field model that can capture these interactions, and handle *real-valued* ratings

Example

FACTORS



Source

Article

Review

User

Example

Instantiation

FACTORS

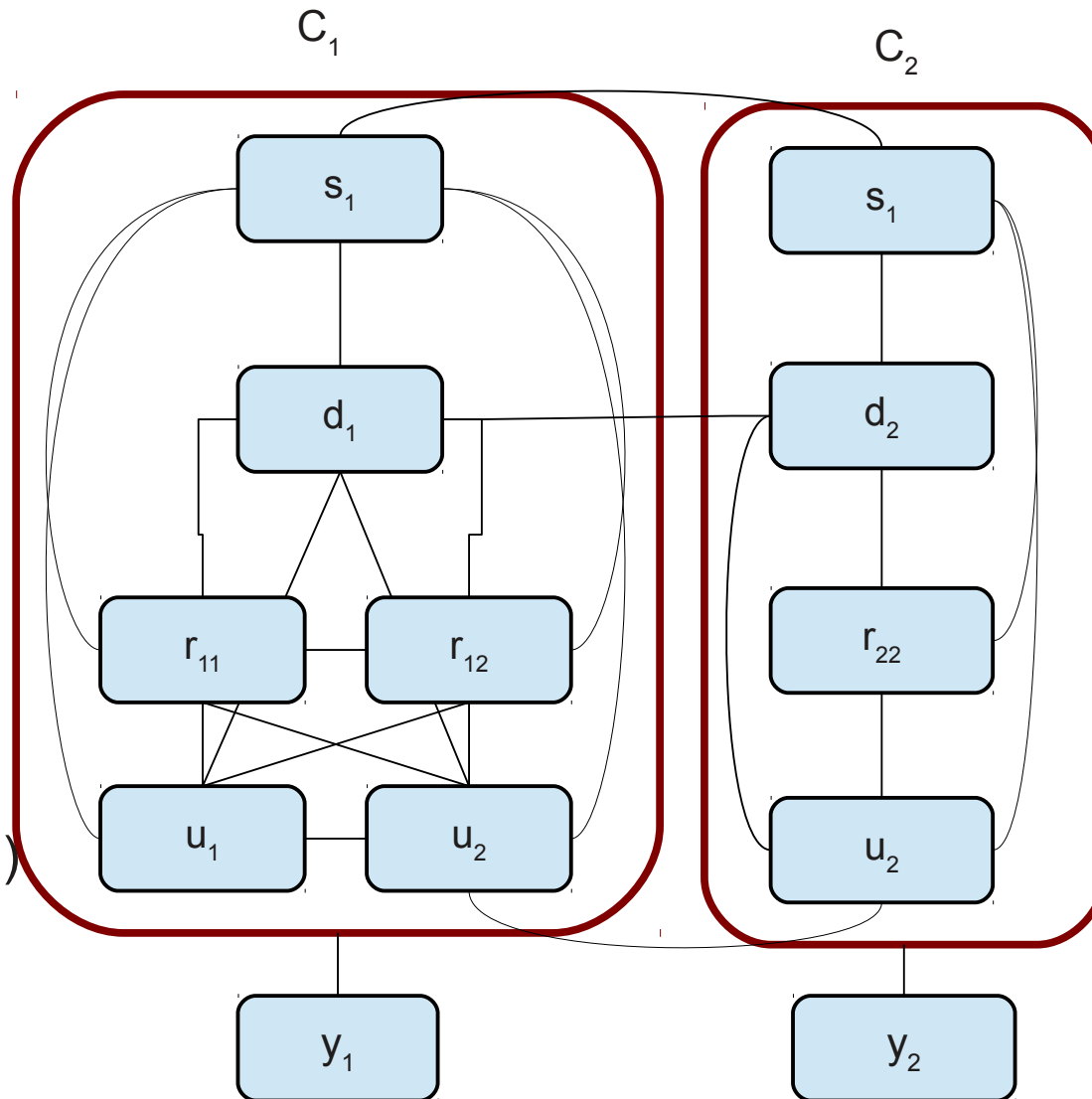
Altnet.org
(progressive/liberal)

Why do conservaties
hate your children?

Topic: Climate

Ratings

Discussions
(liberal vs.conservative)



Source

Article

Review

User

Example

FEATURES

Viewpoint, Expertise

Why do conservatives hate your children?

Topic: Climate

Ratings

Discussions
(liberal vs. conservative)

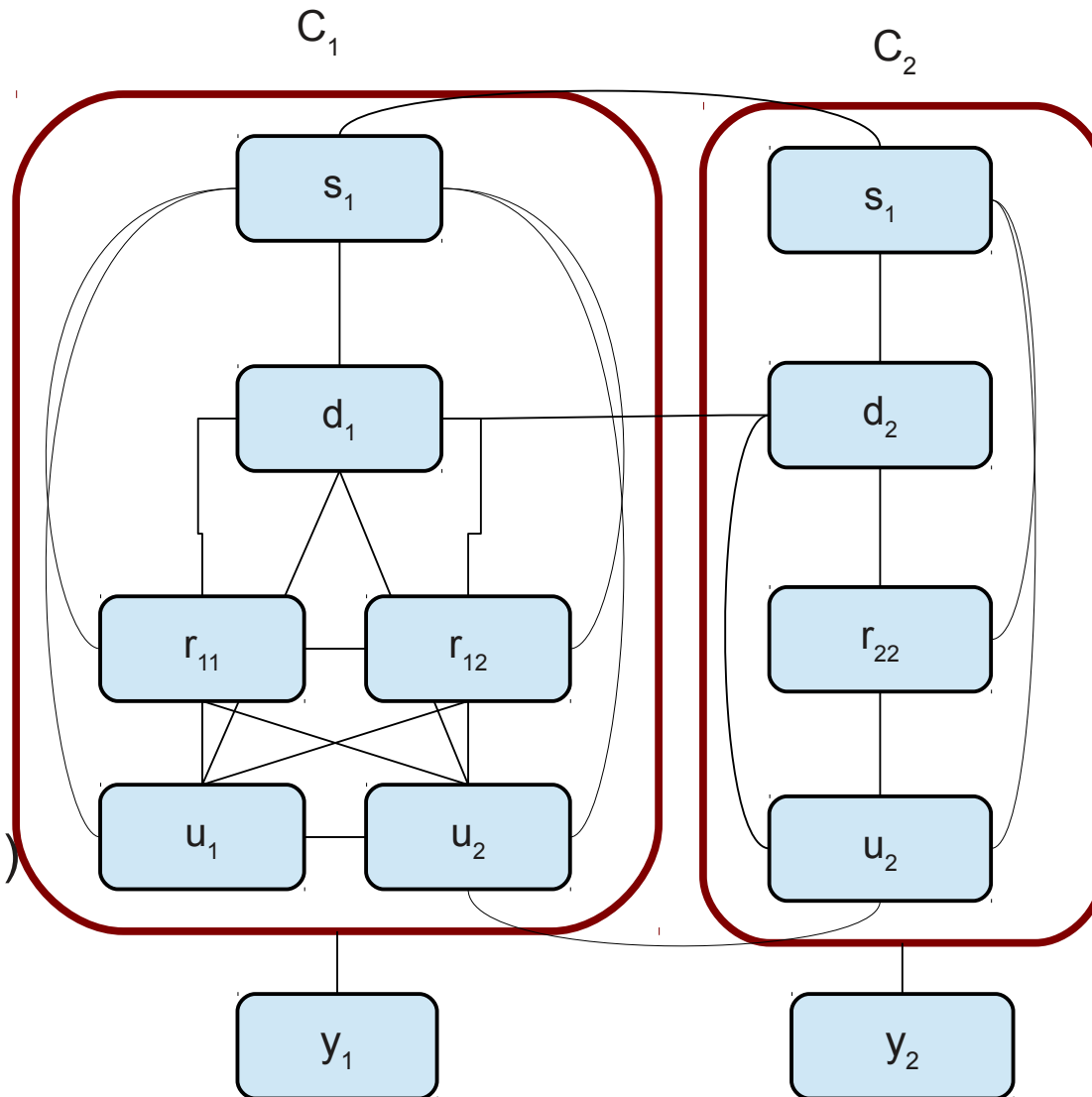
FACTORS

Source

Article

Review

User



Example

FEATURES

FACTORS

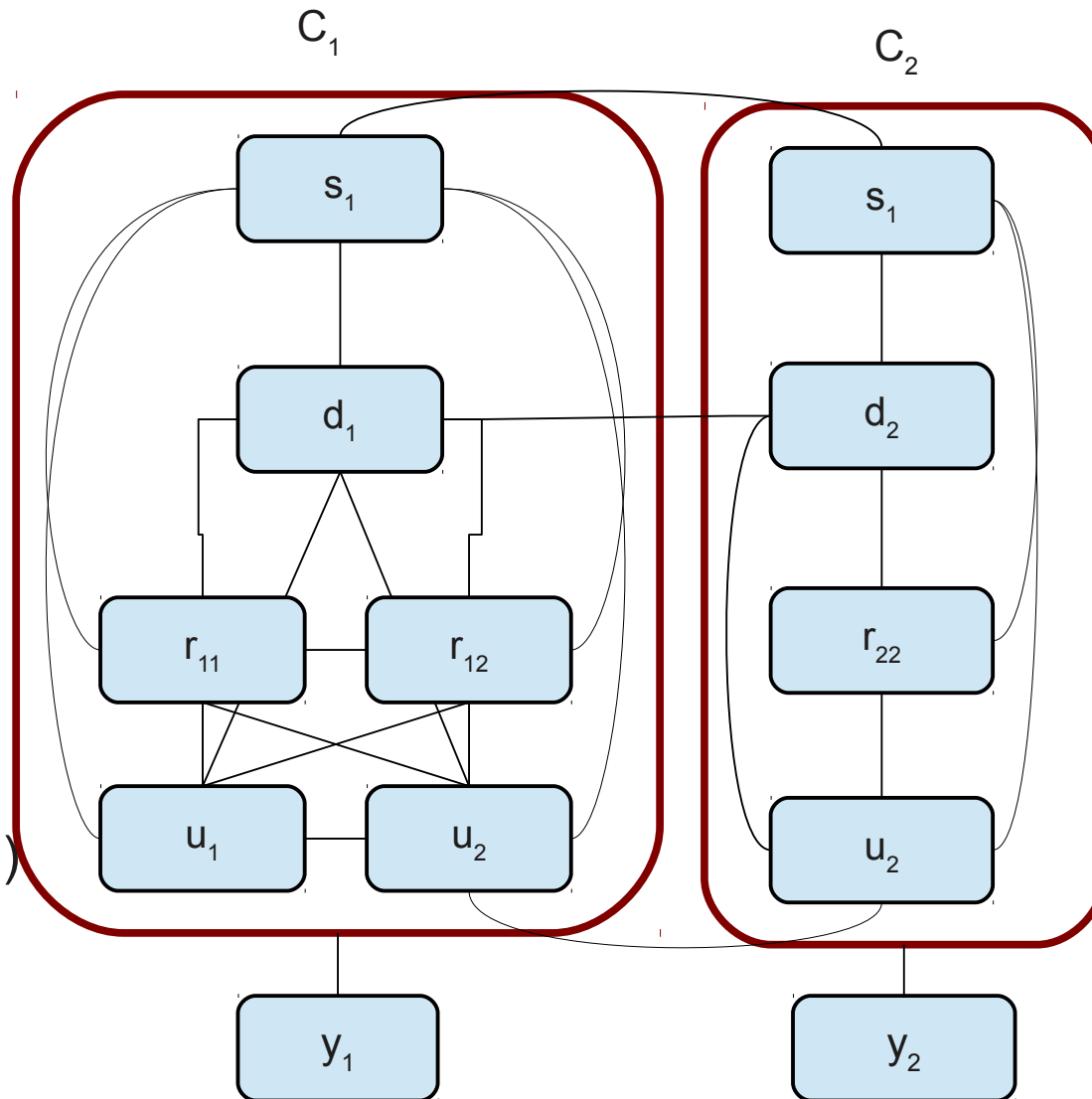
Viewpoint, Expertise

Emotionality, Discourse

Topic: Climate

Ratings

Discussions
(liberal vs. conservative)



Source

Article

Review

User

y_1

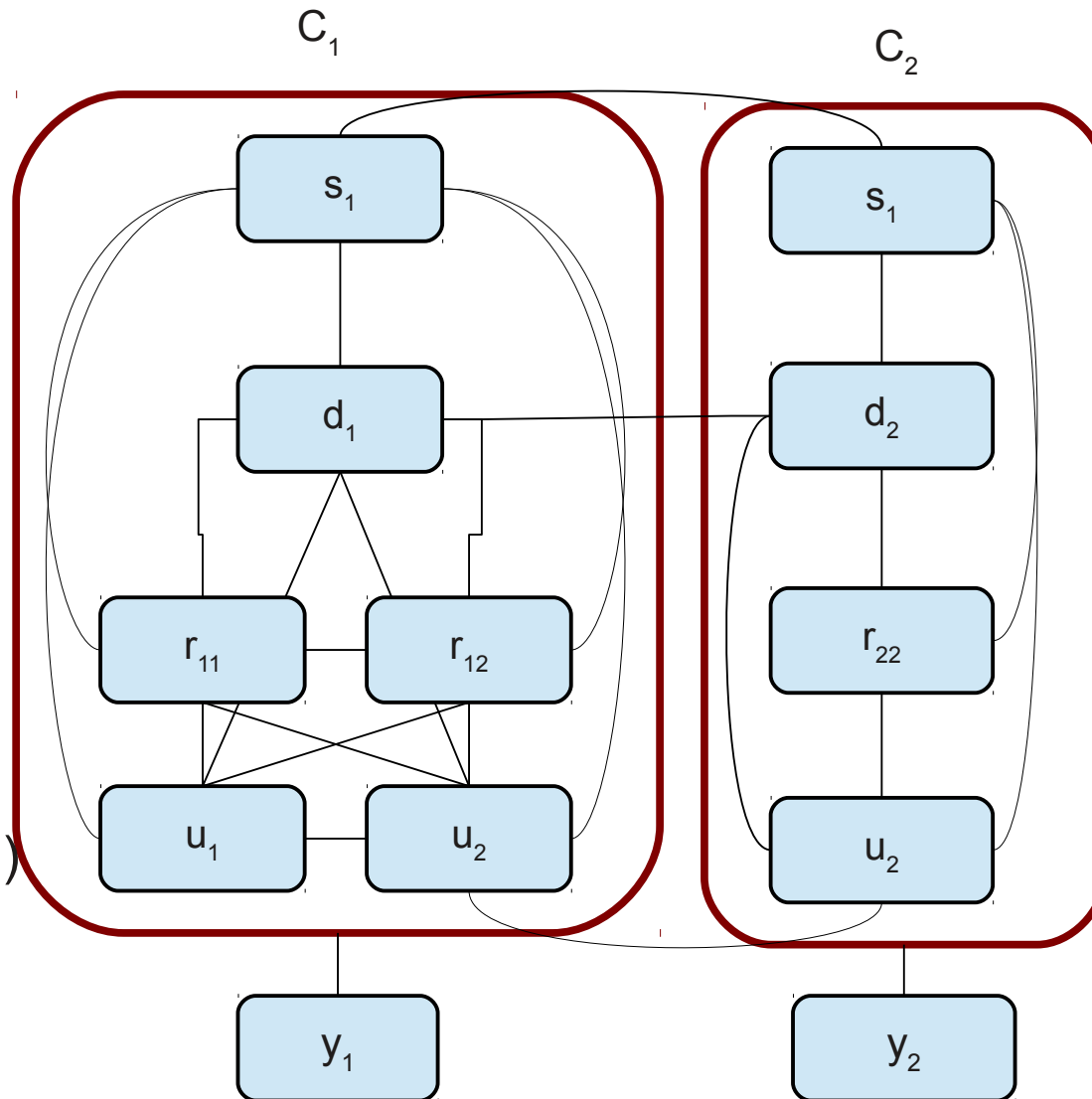
y_2

Example

FEATURES

FACTORS

Viewpoint, Expertise
Emotionality, Discourse
Topic
Ratings
Discussions
(liberal vs. conservative)



Source

Article

Review

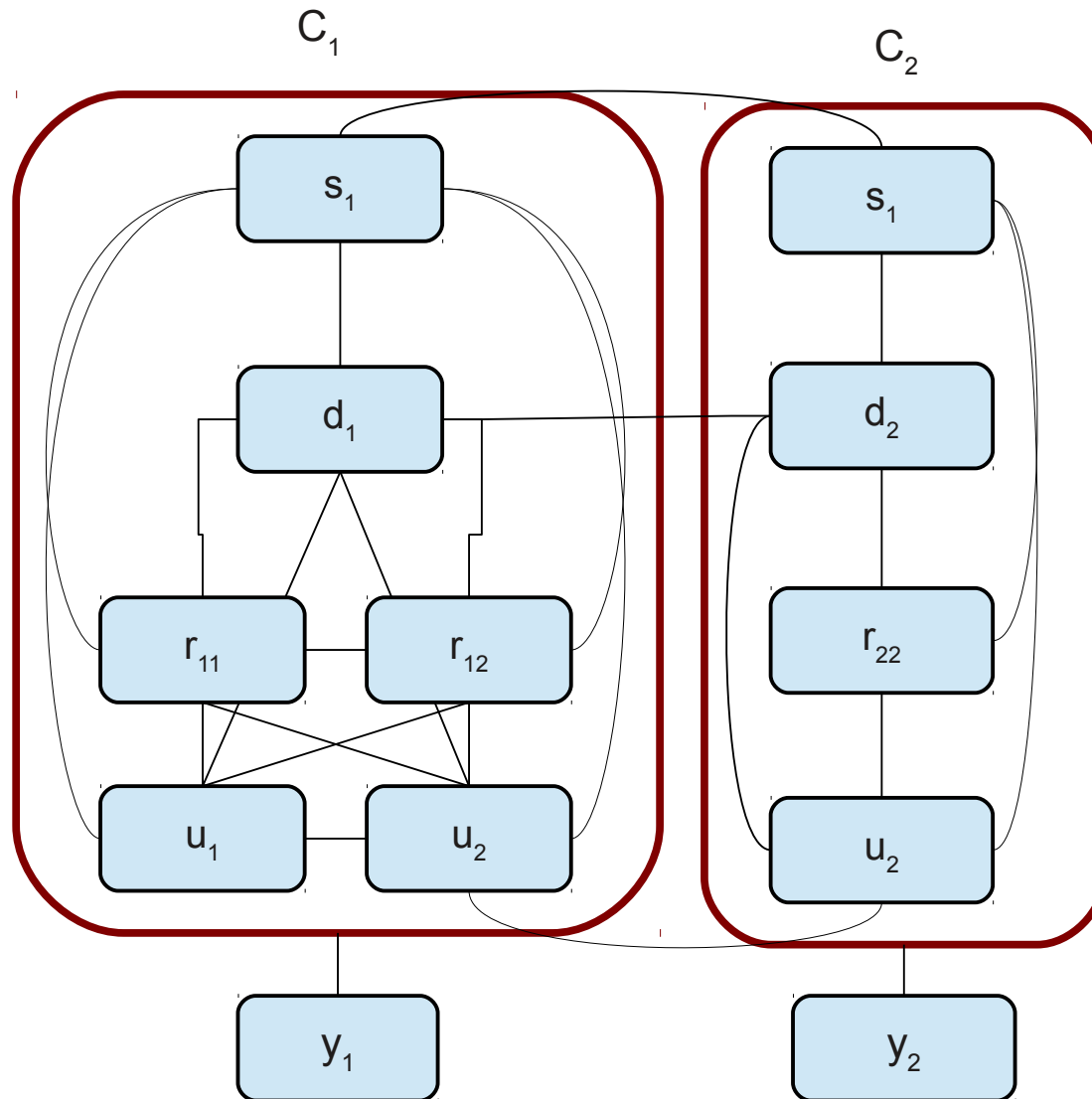
User

Example

FEATURES

FACTORS

Viewpoint, Expertise
Emotionality, Discourse
Topic
Ratings
Bias, Viewpoint,
Expertise



Source

Article

Review

User

Task

ATTRIBUTES

FACTORS

Trustworthiness

Objectivity

Credibility

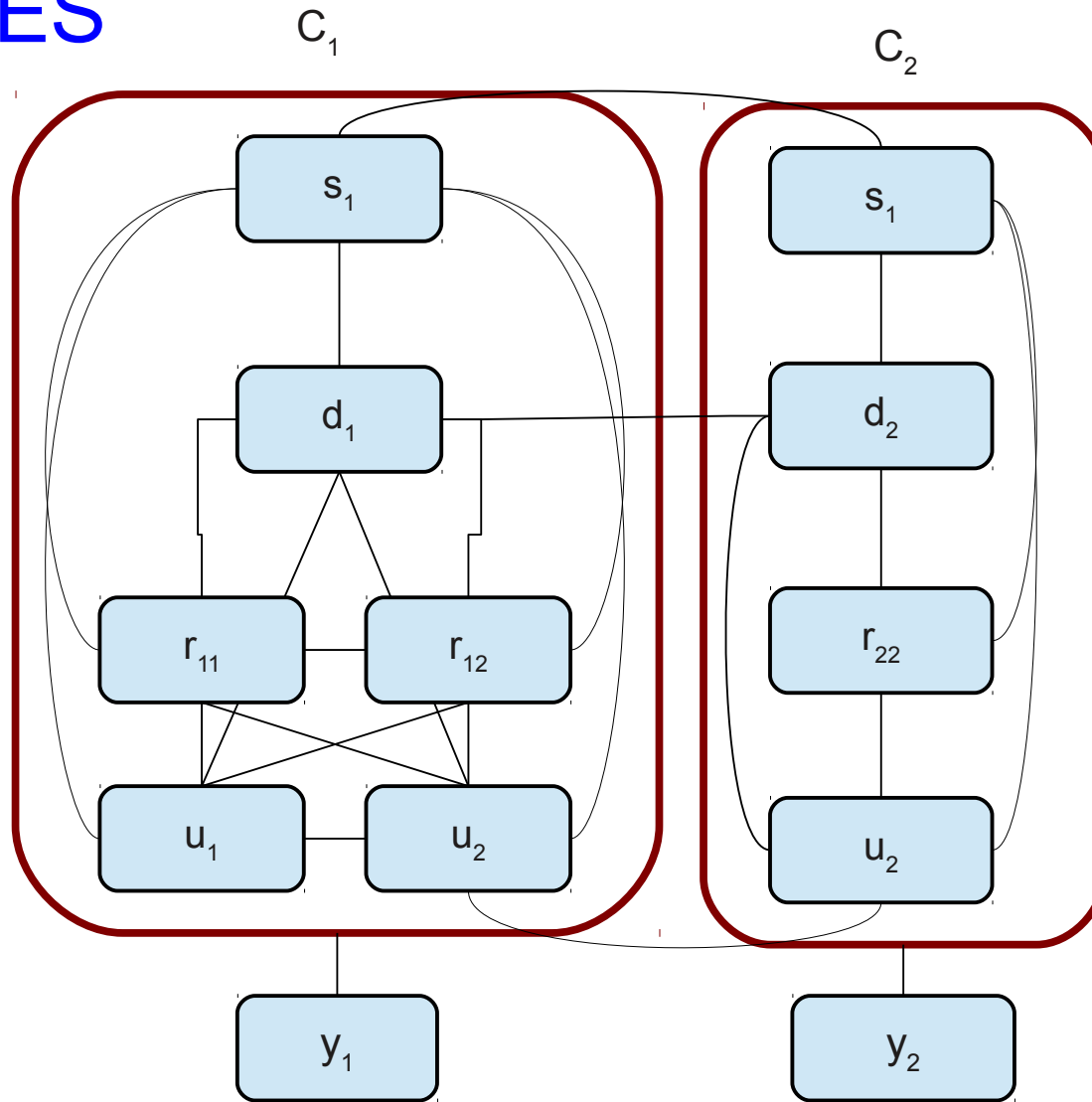
Expertise

Source

Article

Review

User

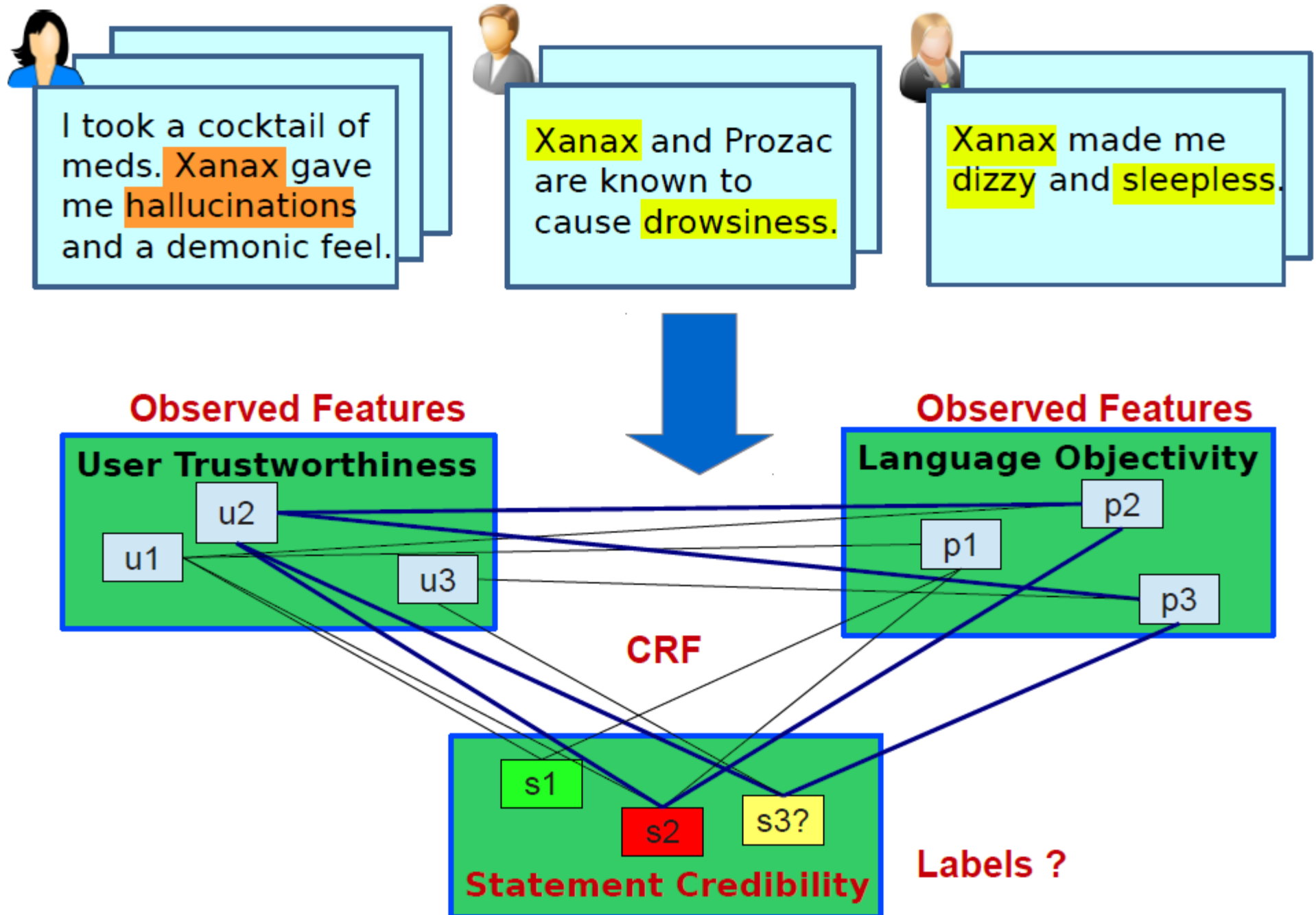


Credibility Analysis

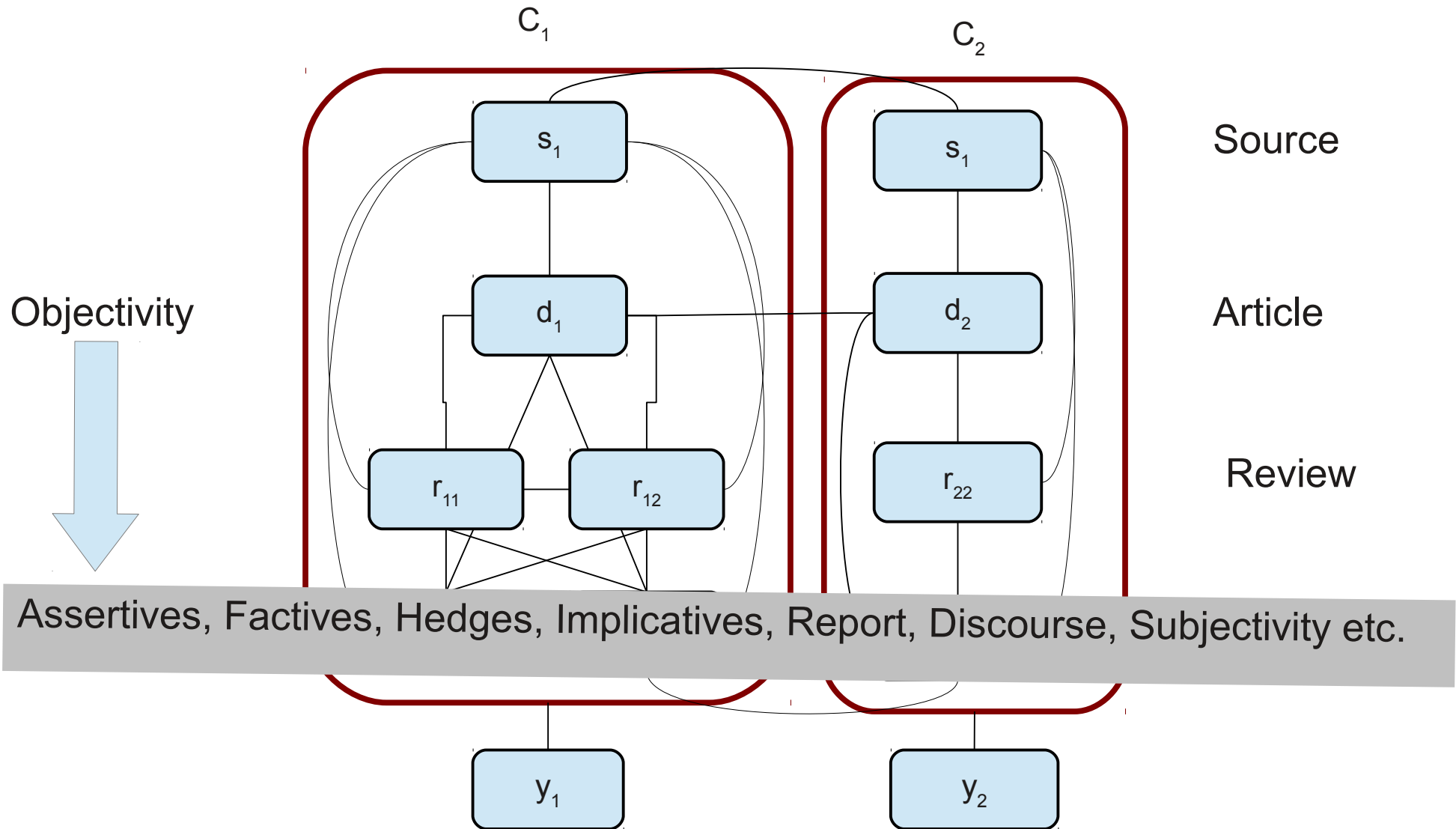
- Given a set of **news sources** generating **news articles**, and **users** reviewing them on different qualitative aspects with **mutual interactions**:
 - Jointly rank the *sources*, *articles*, and *users* based on their *trustworthiness*, *credibility*, and *expertise*

Credibility of Statements in Health Communities

[S. Mukherjee et al.: KDD'14]



Language Features



1. M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. Linguistic models for analyzing and detecting biased language. In ACL, 2013.

2. S. Mukherjee, G. Weikum, and C. Danescu-Niculescu-Mizil. People on drugs: Credibility of user statements in health communities. KDD, 2014.

Topic Features

Latent Topics	Topic Words
Obama admin.	obama, republican, party, election, president, senate, gop, vote
Citizen journ.	cjr, journalism, writers, cjrs, marx, hutchins, reporting, liberty, guides
US military	iraq, war, military, iran, china, nuclear, obama, russia, weapons
AmyGoodman	democracy, military, civil, activist, protests, killing, navajo, amanda
Alternet	media, politics, world news, activism, world, civil, visions, economy
Climate	energy, climate, power, water, change, global, nuclear, fuel, warming

- Only 33% of the articles have *explicit* tags
- Use Latent Dirichlet Allocation to learn the *latent* topic distribution in the corpus of news articles

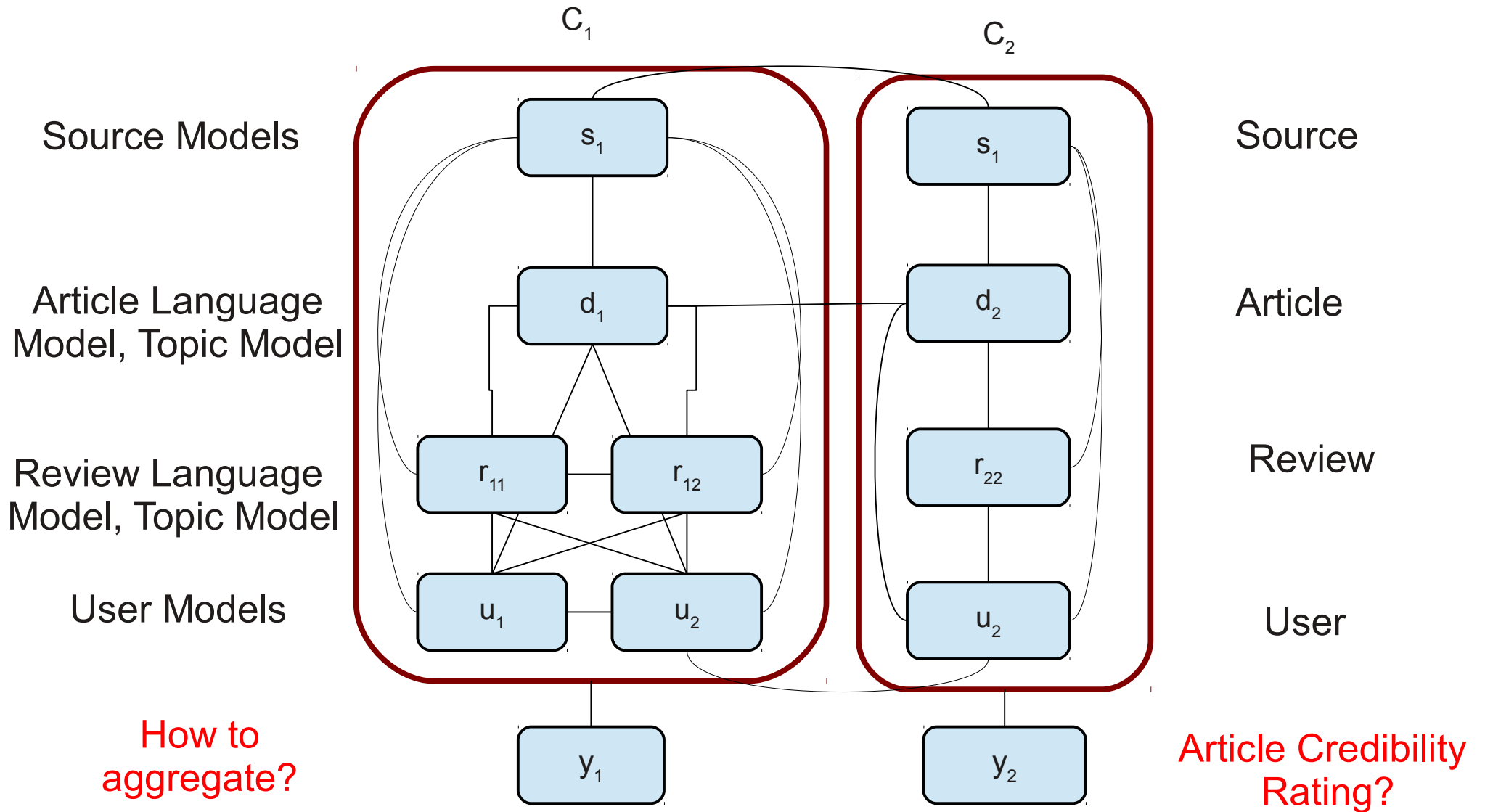
Source Features

Category	Elements
Media	newspaper, blog, radio, magazine, online
Format	editorial, investigative report, news, research
Scope	local, state, regional, national, international
Viewpoint	far left, left, center, right, neutral
Top Topics	politics, weather, war, science,, U.S. military
Expertise on Topics	U.S. congress, Middle East, crime, presidential election, Bush administration, global warming

User Features

Category	Elements
Engagement	answers, ratings (given / received), comments etc.
Agreement	Inter-user agreement
Topics	perspective and expertise
Interactions	user-user, user-item, user-source

Given a factor, with its features, use Support Vector Regression to learn a model that will predict its rating for an article.



Conditional Random Field

Probability Mass Function for discrete labels:

$$p(y|X) = \frac{\exp(\Psi)}{\sum_y \exp(\Psi)}$$

Probability Density Function for continuous ratings:

$$p(y|X) = \frac{\exp(\Psi)}{\int_{-\infty}^{\infty} \exp(\Psi) dy}$$

Energy Function

Clique potential

User Potential

$$\psi(y, s, d, \langle u \rangle, \langle r \rangle) = - \sum_u \alpha_u \mathbb{I}_u(d) (y - \text{SVR}_u)^2$$

$$- \sum_s \beta_s \mathbb{I}_s(d) (y - \text{SVR}_s)^2 - \gamma_1 (y - \text{SVR}_L)^2 - \gamma_2 (y - \text{SVR}_T)^2$$

Source Potential

Language Potential

Topic Potential



Clique: source, article, <users>, <reviews>

partitions the user space

user expertise

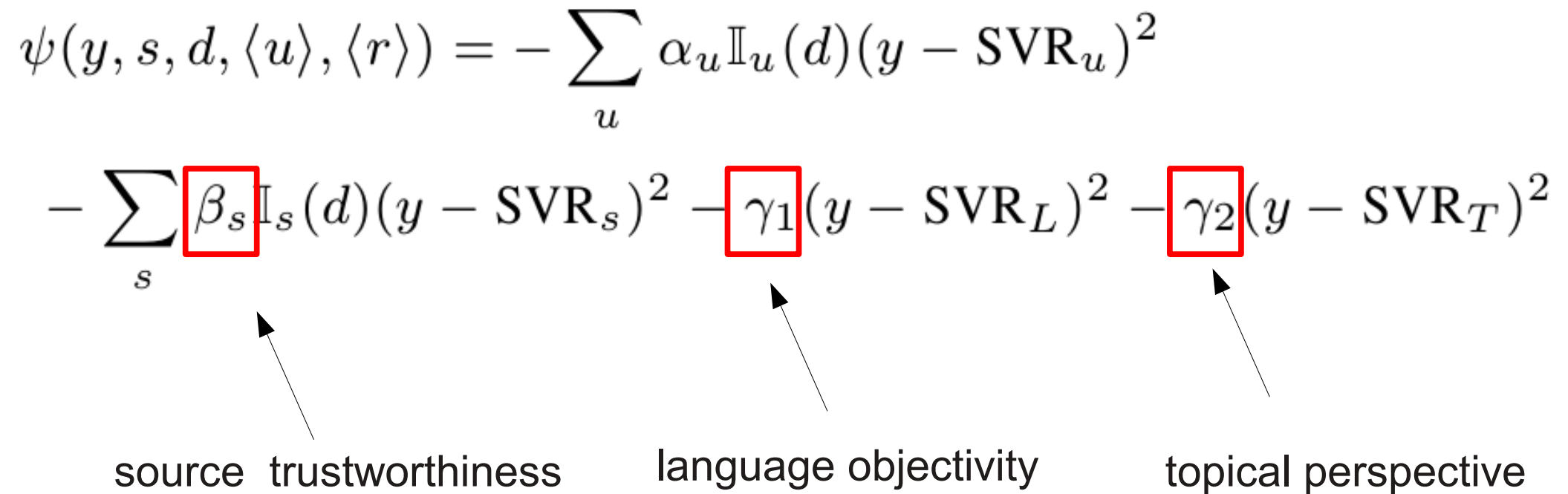
error of predictor SVR

$$\psi(y, s, d, \langle u \rangle, \langle r \rangle) = - \sum_u \alpha_u \mathbb{I}_u(d) (y - \text{SVR}_u)^2 - \sum_s \beta_s \mathbb{I}_s(d) (y - \text{SVR}_s)^2 - \gamma_1 (y - \text{SVR}_L)^2 - \gamma_2 (y - \text{SVR}_T)^2$$

Energy Function

$$\psi(y, s, d, \langle u \rangle, \langle r \rangle) = - \sum_u \alpha_u \mathbb{I}_u(d) (y - \text{SVR}_u)^2$$
$$- \sum_s \beta_s \mathbb{I}_s(d) (y - \text{SVR}_s)^2 - \gamma_1 (y - \text{SVR}_L)^2 - \gamma_2 (y - \text{SVR}_T)^2$$

source trustworthiness language objectivity topical perspective



$$P(y|X) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(y - \mu)^T \Sigma^{-1}(y - \mu)\right)$$

The joint p.d.f is a multivariate gaussian distribution

Σ needs to be positive definite for inverse to exist \rightarrow
 $\{\alpha, \beta, \gamma\} > 0$

Makes sense: predictor reliability should be *positive*

Constrained optimization problem.
Gradient ascent cannot be directly used.

Maximize log-likelihood with respect to $\log \lambda_k$
instead of λ_k

Prediction is the expected value of the function given by
the mean of the Multivariate Gaussian distribution:

$$y' = \operatorname{argmax}_y P(y|X) = \mu = \Sigma b$$

Experiments: NewsTrust

Factors	Count
Unique news articles reviewed in NewsTrust	62,064
NewsTrust stories on news articles	84,704
NewsTrust stories with ≥ 1 reviews	43,107
NewsTrust stories with > 3 reviews	18,521
NewsTrust member reviews of news articles	134,407
News articles extracted from original sources	47,565
NewsTrust stories on extracted news articles	52,579
News sources	5,658
Journalists who wrote news articles	19,236
Timestamps (month and year) of posted news articles	3,122
NewsTrust members who reviewed news articles	7,114
NewsTrust members who posted news articles	1,580
News sources reviewed by NewsTrust members	668
Explicit topic tags	456
Latent topics extracted	300

Data available at: <http://www.mpi-inf.mpg.de/impact/credibilityanalysis/>

Predicting User Ratings

	Model	MSE
	Latent Factor Models (LFM)	
Users, Articles, Ratings	Simple LFM	0.95
+Time	Experience-based LFM	0.85
+Review Text	Text-based LFM	0.78
+Review Text and Interactions	Our Model: User SVR	0.60

1. Y. Koren. Factorization meets the neighborhood: A multifaceted collaborative filtering model. KDD, 2008.
2. J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. RecSys, 2013.
3. J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In WWW, 2013.

Predicting Article Credibility Ratings

Model	Only Title MSE	Title & Text MSE
Language Model: SVR		
Language (Bias and Subjectivity)	3.89	0.72
Explicit Topics	1.74	1.74
Explicit + Latent Topics	1.68	1.01
All Topics (Explicit + Latent) + Language	1.57	0.61

Predicting Article Credibility Ratings

Model	Only Title MSE	Title & Text MSE
Language Model: SVR		
Language (Bias and Subjectivity)	3.89	0.72
Explicit Topics	1.74	1.74
Explicit + Latent Topics	1.68	1.01
All Topics (Explicit + Latent) + Language	1.57	0.61
News Source Features and Language Model: SVR		
News Source	1.69	1.69
News Source + All Topics + Language	0.91	0.46

Predicting Article Credibility Ratings

Model	Only Title MSE	Title & Text MSE
Language Model: SVR		
Language (Bias and Subjectivity)	3.89	0.72
Explicit Topics	1.74	1.74
Explicit + Latent Topics	1.68	1.01
All Topics (Explicit + Latent) + Language	1.57	0.61
News Source Features and Language Model: SVR		
News Source	1.69	1.69
News Source + All Topics + Language	0.91	0.46
Aggregated Model: SVR		
Users + All Topics + Language + News Source	0.43	0.41

Predicting Article Credibility Ratings

Model	Only Title MSE	Title & Text MSE
Language Model: SVR		
Language (Bias and Subjectivity)	3.89	0.72
Explicit Topics	1.74	1.74
Explicit + Latent Topics	1.68	1.01
All Topics (Explicit + Latent) + Language	1.57	0.61
News Source Features and Language Model: SVR		
News Source	1.69	1.69
News Source + All Topics + Language	0.91	0.46
Aggregated Model: SVR		
Users + All Topics + Language + News Source	0.43	0.41
Our Model: CCRF+SVR		
User + All Topics + Language + News Source	0.36	0.33

Ranking Trustworthy News Sources

Model	NDCG
Experience LFM	0.80
PageRank	0.83
CCRF	0.86

Ranking Expert Users:

Model	NDCG
Experience LFM	0.81
Member Ratings	0.85
CCRF	0.91

Sample Output: Most and Least Trust Sources on Sample Topics

Money - Politics	War in Iraq	Media - Politics	Green Technology
Most Trusted			
rollingstone.com	nybooks.com	consortiumnews	discovermagazine.com
truthdig.com	consortiumnews	thenation.com	nature.com
democracynow.org	truthout.org	thedailyshow.com	scientificamerican.com
Least Trusted			
firedoglake.com	crooksandliars	rushlimbaugh.com	
suntimes.com	timesonline	rightwingnews.com	
trueslant.com	suntimes.com	foxnews.com	

Conclusions

- Joint interaction between *language*, *topics*, *users* and *sources* lead to better prediction in multiple tasks
- User *expertise*, source *trustworthiness*, language *objectivity*, topical *perspective* and article *credibility* mutually reinforce each other

Ongoing Work

- Analyze temporal evolution of these factors
- Communities are inherently dynamic in nature
- Source trustworthiness, and user expertise change with time
- To this end we propose an Experience-aware Item Recommendation for Evolving Review Communities, ICDM 2015.