Graph-Based Lexicon Expansion with Sparsity-Inducing Penalties

Dipanjan Das, LTI, CMU $\rightarrow$ Google
Noah Smith, LTI, CMU

Thanks: André Martins, Amar Subramanya, and Partha Talukdar. This research was supported by Qatar National Research Foundation grant NPRP 08-485-1-083, Google, and TeraGrid resources provided by the Pittsburgh Supercomputing Center under NSF grant number TG-DBS110003.

Motivation

- FrameNet lexicon (Fillmore et al., 2003)
  - For many words, a set of abstract semantic frames
  - E.g., contribute/V can evoke GIVING or SYMPTOM
- SEMAFOR (Das et al., 2010).
  - Finds: frames evoked + semantic roles

What about the words not in the lexicon or data?

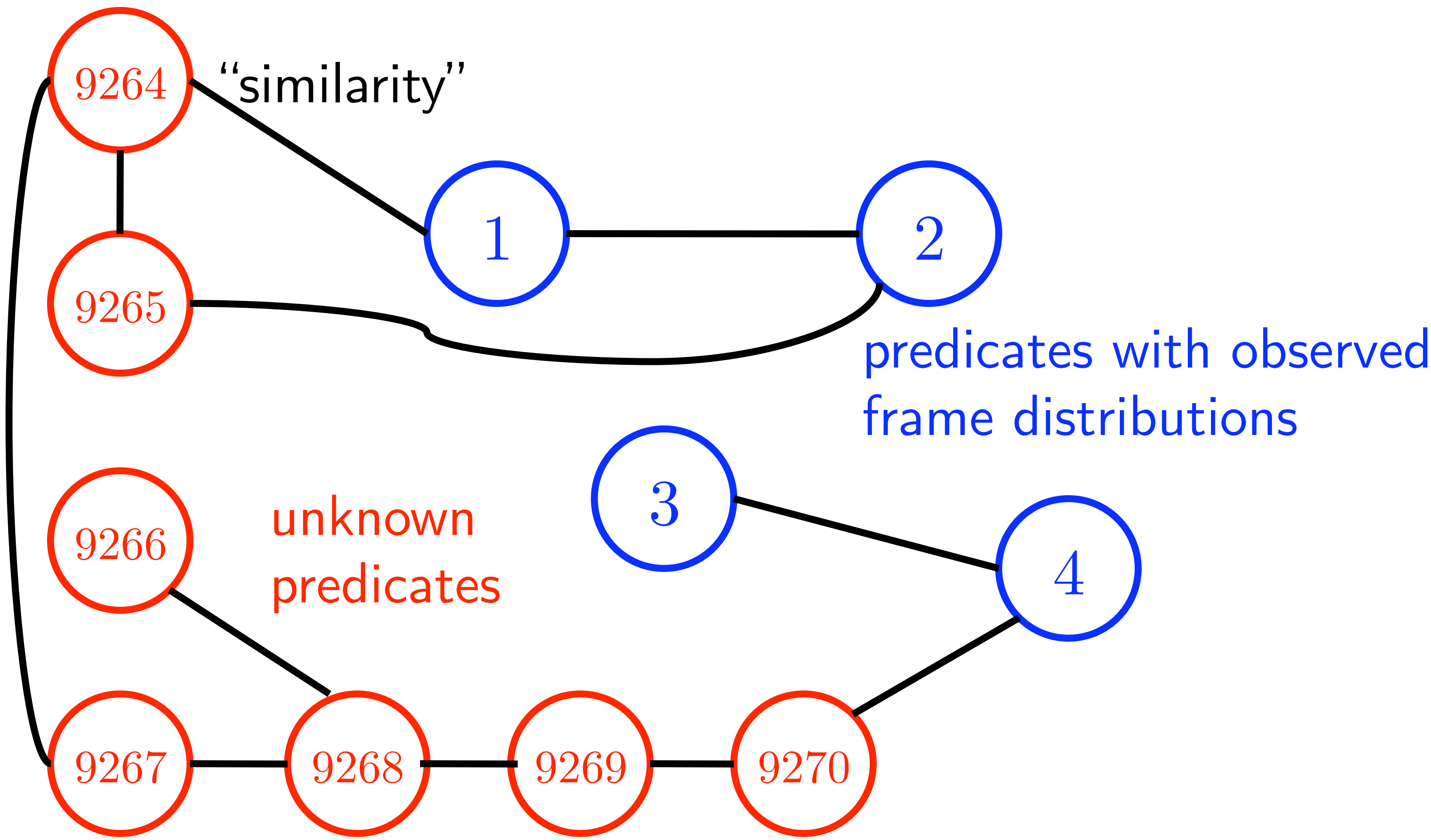
Das and Smith (2011)

- **Graph-based semi-supervised learning** with quadratic penalties (Bengio et al., 2006; Subramanya et al., 2010).
  - Frame identification F_1 on unknown predicates: $47\% \rightarrow 62\%$
  - Frame parsing F_1 on unknown predicates: $30\% \rightarrow 44\%$

Das and Smith (2011)

- **Graph-based semi-supervised learning** with quadratic penalties (Bengio et al., 2006; Subramanya et al., 2010).
  - Frame identification F_1 on unknown predicates: 47% $\rightarrow$ 62% $\rightarrow$ (today) 65%
  - Frame parsing F_1 on unknown predicates: 30% $\rightarrow$ 44% $\rightarrow$ (today) 47%
- Today: we consider alternatives that target *sparsity*, or each word associating with relatively few frames.

Graph-Based Learning



The Case for Sparsity

- Lexical ambiguity is pervasive, but each word's ambiguity is fairly limited.
- Ruling out possibilities $\rightarrow$ better runtime and memory properties.

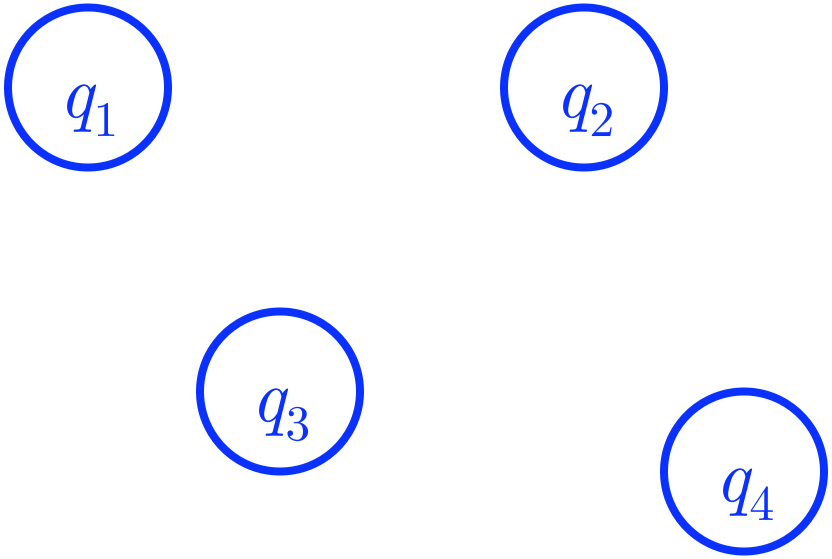
Outline

1. A general family of graph-based SSL techniques for learning distributions.
   - Defining the graph
   - Constructing the graph and carrying out inference
   - New: sparse and unnormalized distributions
2. Experiments with frame analysis: favorable comparison to state-of-the-art graph-based learning algorithms

Notation

- $T =$ the set of types (words)
- $L =$ the set of labels (frames)
- Let $q_t(l)$ denote the estimated probability that type t will take label l.

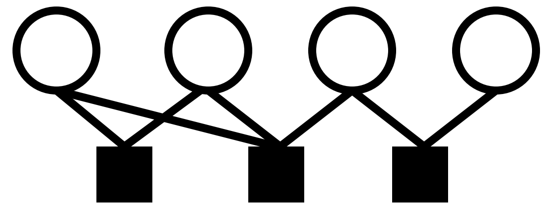
Vertices, Part 1



Think of this as a **graphical model** whose random variables take **vector** values.

Factor Graphs (Kschischang et al., 2001)

- Bipartite graph:
  - Random variable vertices V
  - "Factor" vertices F



- Distribution over all variables' values:

$$\log P\left(\{v\}_{v\in V}\right) \quad = \quad -\log Z + \sum_{f\in F} \log \alpha_f \left(\{v\}_{(v,f)\in E}\right)$$

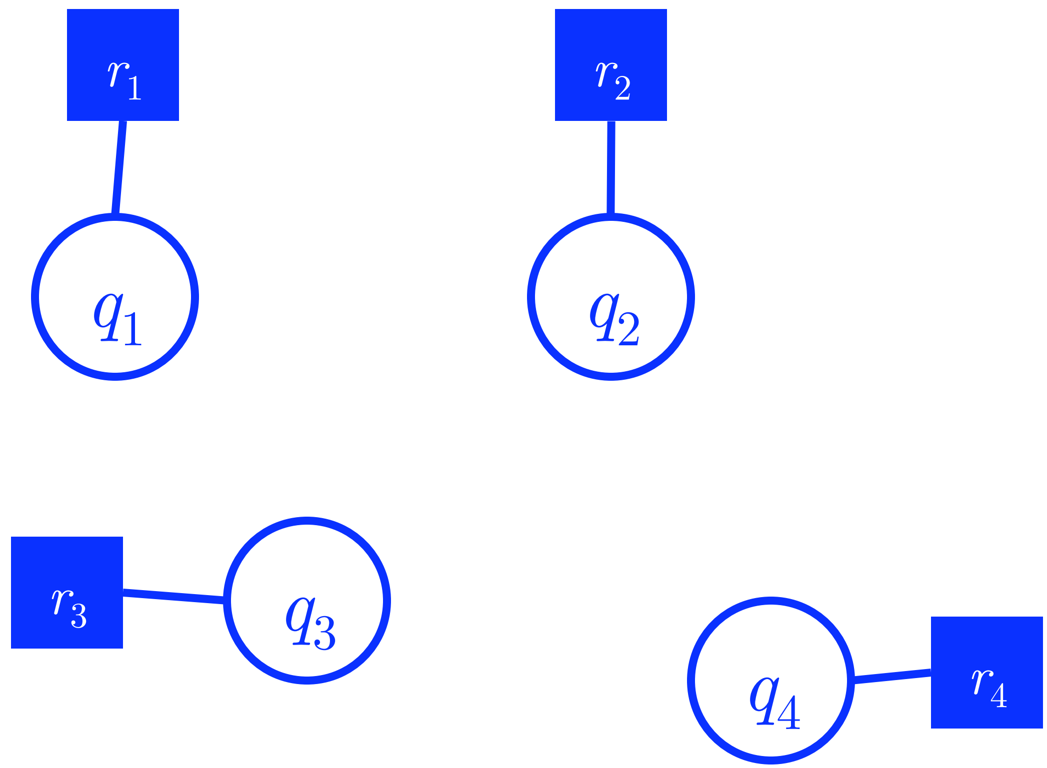
- Today: finding collectively highest-scoring values (MAP inference) ≡ estimating q
  - Log-factors ≡ negated penalties

Notation

- $T =$ the set of types (words)
- $L =$ the set of labels (frames)
- Let $q_t(l)$ denote the estimated probability that type t will take label l.

- Let $r_t(l)$ denote the observed relative frequency of type t with label l.

Penalties (1 of 3)



"Each type t_i's value should be close to its empirical distribution r_i."

Empirical Penalties

- "Gaussian" (Zhu et al., 2003): penalty is the squared L_2 norm

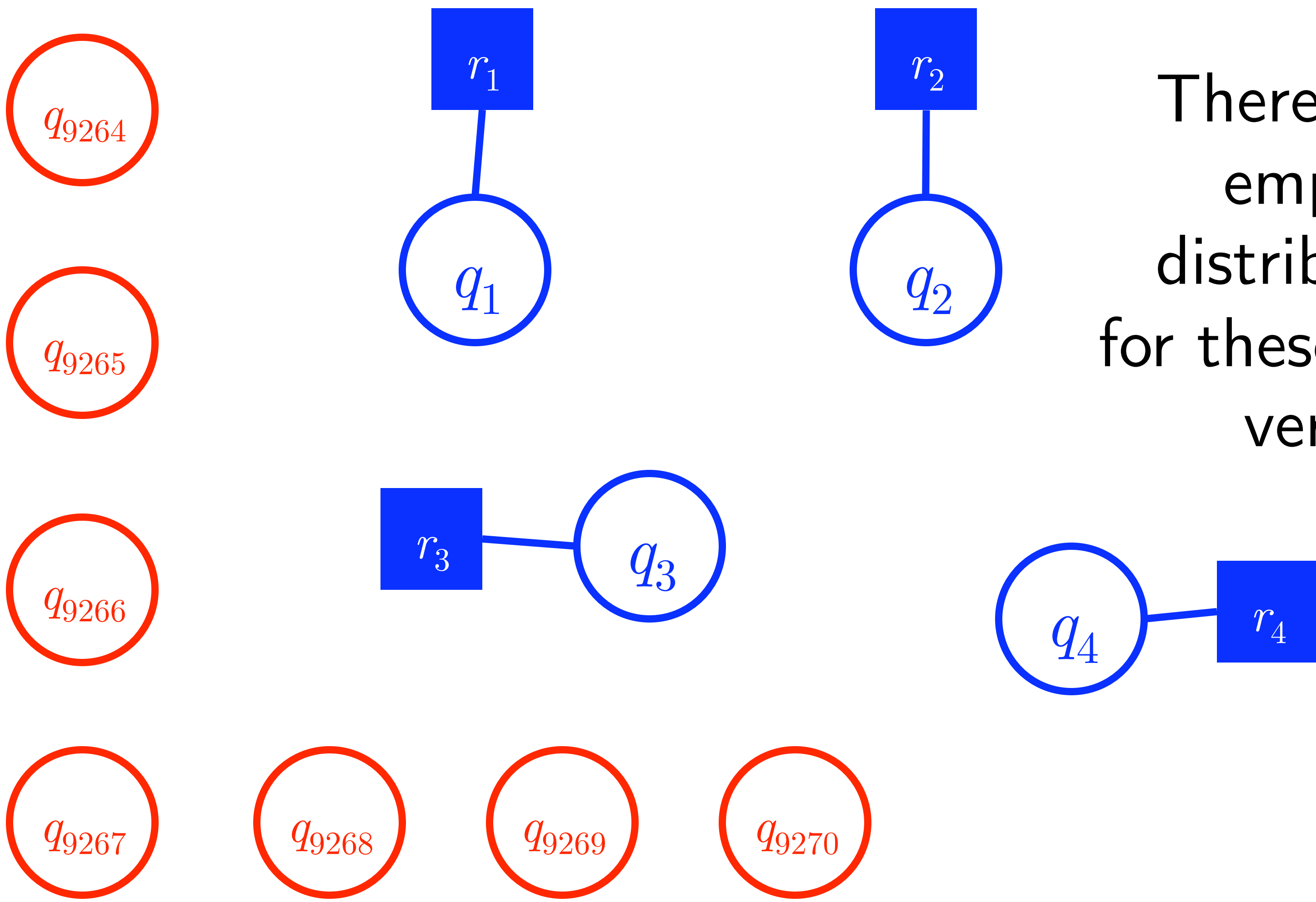
$$\log \phi_t \left(q_t, r_t\right) \quad = \quad -\|q_t - r_t\|_2^2$$

- "Entropic": penalty is the JS-divergence (cf. Subramanya and Bilmes, 2008, who used KL)

$$\log \phi_t \left(q_t, r_t\right) \quad = \quad -\frac{1}{2}\left(D\left(q_t \middle\| \frac{q_t + r_t}{2}\right) + D\left(r_t \middle\| \frac{q_t + r_t}{2}\right)\right)$$

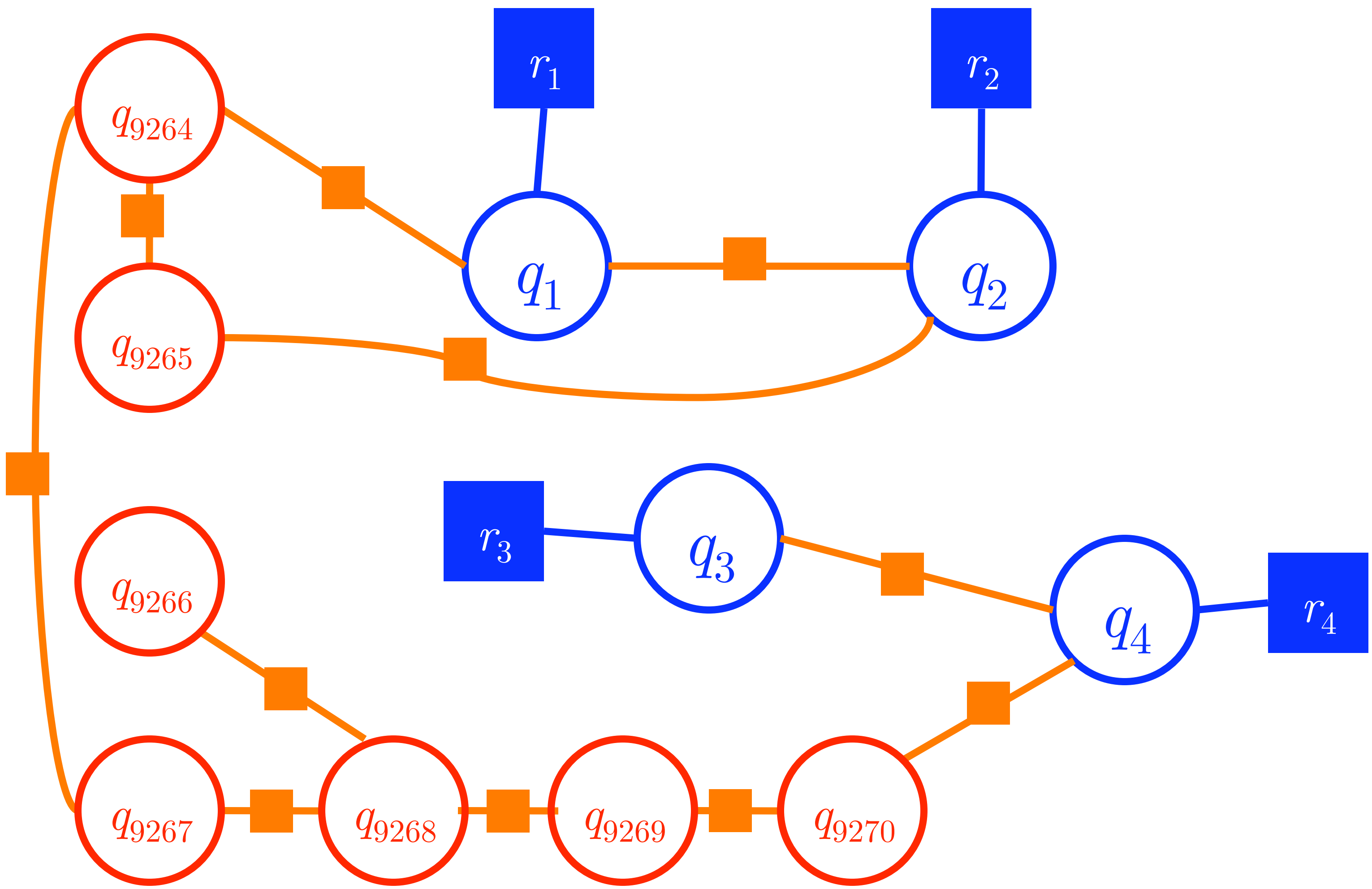
Let's Get Semi-Supervised

Vertices, Part 2



There is no empirical distribution for these new vertices!

Penalties (2 of 3)



Similarity Factors

$$\log \varphi_{t,t'}(q_t, q_{t'}) = -2 \cdot \mu \cdot \mathsf{sim}(t, t') \cdot \|q_t - q_{t'}\|_2^2$$

"Gaussian"

$$\log \varphi_{t,t'}(q_t, q_{t'}) = -2 \cdot \mu \cdot \mathsf{sim}(t, t') \cdot \mathrm{JS}(q_t \| q_{t'})$$

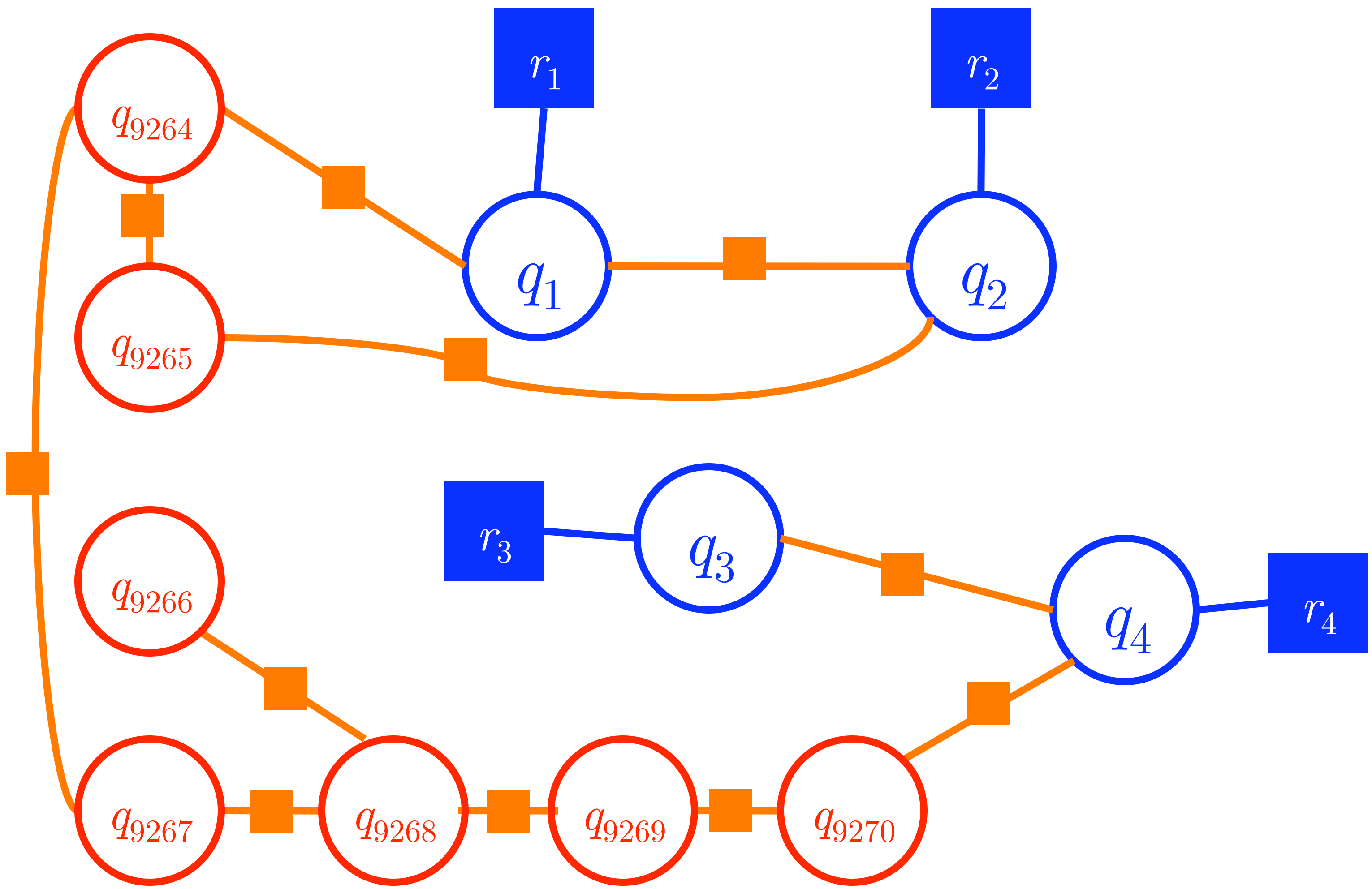
"Entropic"

Constructing the Graph

in one slide

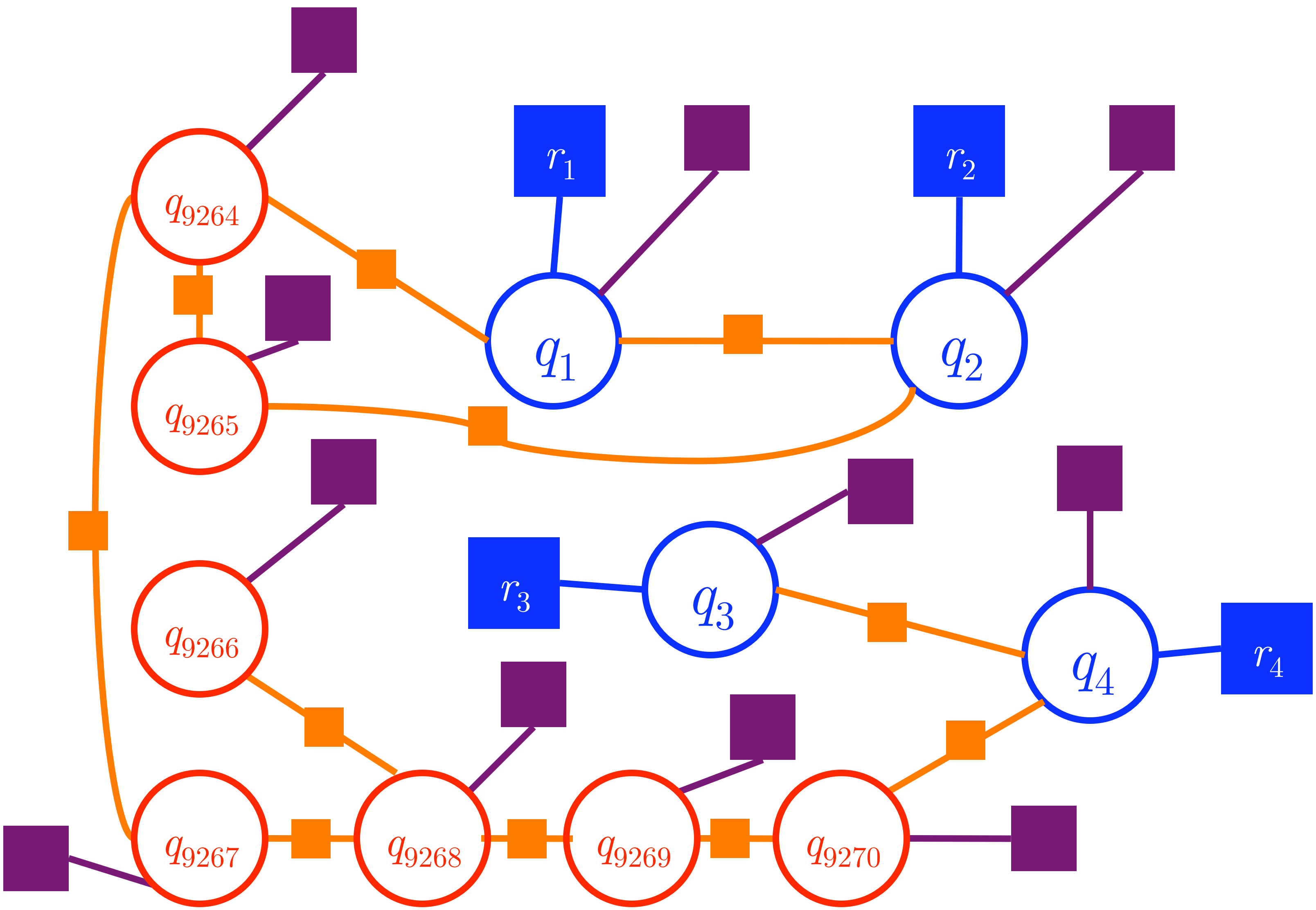
- Conjecture: *contextual* distributional similarity correlates with *lexical* distributional similarity.
  - Subramanya et al. (2010); Das and Petrov (2011); Das and Smith (2011)

1. Calculate distributional similarity for each pair.
   - Details in past work; nothing new here.
2. Choose each vertex's K closest neighbors.
3. Weight each log-factor by the similarity score.



r_1
r_2
q_{9264}
q_1
q_2
q_{9265}
r_3
q_3
q_4
r_4
q_{9266}
q_{9267}
q_{9268}
q_{9269}
q_{9270}

Penalties (3 of 3)



What Might
Unary Penalties/Factors Do?

- Hard factors to enforce nonnegativity, normalization
- Encourage near-uniformity
  - squared distance to uniform (Zhu et al., 2003; Subramanya et al., 2010; Das and Smith, 2011)
  - entropy (Subramanya and Bilmes, 2008)
- Encourage sparsity
  - *Main goal of this paper!*

Unary Log-Factors

$$\log \psi_t(q_t) \quad =$$

- Squared distance to uniform: $-\lambda \left\| q_t - \frac{1}{|L|} \right\|_2^2$

- Entropy: $\lambda H(q_t)$

- "Lasso"/L_1 (Tibshirani, 1996): $-\lambda \left\| q_t \right\|_1$

- "Elitist Lasso"/squared $L_{1,2}$ (Kowalski and Torrésani, 2009):

$$-\lambda \left(\left\| q_t \right\|_1 \right)^2$$

Models to Compare

Model	**Empirical and pairwise factors**	**Unary factor**
normalized Gaussian field (Das and Smith, 2011; generalizes Zhu et al., 2003)	Gaussian	squared L_2 to uniform, normalization
"measure propagation" (Subramanya and Bilmes, 2008)	Kullback-Leibler	entropy, normalization
UGF-L_2	Gaussian	squared L_2 to uniform
UGF-L_1	Gaussian	lasso (L_1)
UGF-$L_{1,2}$	Gaussian	elitist lasso (squared $L_{1,2}$)
UJSF-L_2	Jensen-Shannon	squared L_2 to uniform
UJSF-L_1	Jensen-Shannon	lasso (L_1)
UJSF-$L_{1,2}$	Jensen-Shannon	elitist lasso (squared $L_{1,2}$)

sparsity-inducing penalties

unnormalized distributions

Where We Are So Far

- "Factor graph" view of semisupervised graph-based learning.
  - Encompasses familiar Gaussian and entropic approaches.
  - Estimating all q_t equates to MAP inference.

Yet to come:

- Inference algorithm for all q_t
- Experiments

Inference

In One Slide

- All of these problems are convex.
- Past work relied on specialized iterative methods.
- Lack of normalization constraints makes things simpler!
  - Easy quasi-Newton gradient-based method, L-BFGS-B (with nonnegativity "box" constraints)
  - Non-differentiability at 0 causes no problems (assume "right-continuity")
  - KL and JS divergence can be generalized to unnormalized measures

Experiment 1

- (see the paper)

Experiment 2: Semantic Frames

- *Types*: word plus POS
- *Labels*: 877 frames from FrameNet
- *Empirical distributions*: 3,256 sentences from FrameNet 1.5 release
- *Graph*: 64,480 vertices (see D&S 2011)
- *Evaluation*: use induced lexicon to constrain frame analysis of unknown predicates on 2,420 sentence test set.
  1. Label words with frames.
  2. … Then find arguments (semantic roles)

Frame Identification

Model	Unknown predicates, partial match F_1	Lexicon size
supervised (Das et al., 2010)	46.62	
normalized Gaussian (Das & Smith, 2011)	62.35	129K
"measure propagation"	60.07	129K
UGF-L_2	60.81	129K
UGF-L_1	62.85	123K
UGF-$L_{1,2}$	62.85	129K
UJSF-L_2	62.81	128K
UJSF-L_1	62.43	129K
UJSF-$L_{1,2}$	**65.29**	**46K**

Learned Frames (UJSF-$L_{1,2}$)

- discrepancy/N: SIMILARITY, NON-COMMUTATIVE-STATEMENT, NATURAL-FEATURES
- contribution/N: GIVING, COMMERCE-PAY, COMMITMENT, ASSISTANCE, EARNINGS-AND-LOSSES
- print/V: TEXT-CREATION, STATE-OF-ENTITY, DISPERSAL, CONTACTING, READING
- mislead/V: PREVARICATION, EXPERIENCER-OBJ, MANIPULATE-INTO-DOING, REASSURING, EVIDENCE
- abused/A: (Our models can assign $q_t = \mathbf{0}$.)
- maker/N: MANUFACTURING, BUSINESSES, COMMERCE-SCENARIO, SUPPLY, BEING-ACTIVE
- inspire/V: CAUSE-TO-START, SUBJECTIVE-INFLUENCE, OBJECTIVE-INFLUENCE, EXPERIENCER-OBJ, SETTING-FIRE
- failed/A: SUCCESSFUL-ACTION, SUCCESSFULLY-COMMUNICATE-MESSAGE

blue = correct

Frame Parsing (Das, 2012)

Model	**Unknown predicates, partial match F_1**
supervised (Das et al., 2010)	29.20
normalized Gaussian (Das & Smith, 2011)	42.71
"measure propagation"	41.41
UGF-L_2	41.97
UGF-L_1	42.58
UGF-$L_{1,2}$	42.58
UJSF-L_2	43.91
UJSF-L_1	42.29
UJSF-$L_{1,2}$	**46.75**

Example

REASON Action

Discrepancies between North Korean declarations and IAEA inspection findings indicate that North Korea might have reprocessed enough plutonium for one or two nuclear weapons.

Example

SIMILARITY Entities

Discrepancies between North Korean declarations and IAEA inspection findings indicate that North Korea might have reprocessed enough plutonium for one or two nuclear weapons.

SEMAFOR

http://www.ark.cs.cmu.edu/SEMAFOR

- Current version (2.1) incorporates the expanded lexicon.
- To hear about algorithmic advances in SEMAFOR, see our *SEM talk, 2pm Friday.

Conclusions

- General family of graph-based semi-supervised learning objectives.
- Key technical ideas:
  - Don't require normalized measures
  - Encourage (local) sparsity
  - Use general optimization methods

Thanks!