



Robust Entity Clustering via Phylogenetic Inference

Nicholas Andrews
with Jason Eisner and Mark Dredze



human language technology
center of excellence

Some data



Did **Taylor swift** just dis harry sytles on the
#grammys ? Lmao

Some data



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys

Some data



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.

Some data



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.



Taylor swift is apart of the Illuminati #grammys

Some data



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.



Taylor swift is apart of the Illuminati #grammys



Ladies STILL love **LL Cool James**.

Some data



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.



Taylor swift is apart of the Illuminati #grammys

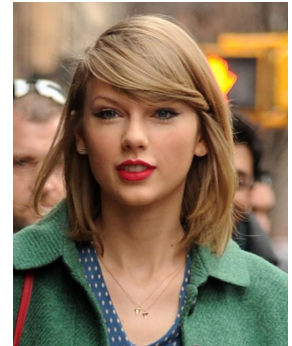


Ladies STILL love **LL Cool James**.



LL Cool J is looking realllll chizzled!

Entity clustering



Did **Taylor swift(1)** just dis harry sytles on the**(1)**
#grammys ? Lmao



Lets see how bad **T Swift(1)** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle**
(1) is on drugs. Lots of drugs.



Taylor swift(1) is apart of the Illuminati
#grammys



Ladies STILL love **LL Cool James(2)**.



LL Cool J(2) is looking reaIIlll chizzled!



(2)

Key idea: “Directed” name variation



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao

Abbreviation



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.



Taylor swift is apart of the Illuminati #grammys



Ladies STILL love **LL Cool James**.



LL Cool J is looking reaIIlll chizzled!

“Directed” name variation



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.



Taylor swift is apart of the Illuminati #grammys



Ladies STILL love **LL Cool James**.



LL Cool J is looking reaIIII chizzled!

“Directed” name variation



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.



Taylor swift is apart of the Illuminati #grammys



Ladies STILL love **LL Cool James**.



LL Cool J is looking reaIIII chizzled!

“Directed” name variation



Did **Taylor swift** just dis harry sytles on the #grammys ? Lmao



Lets see how bad **T Swift** will be. #grammys



Watching the Grammy's - it's clear that **T-Swizzle** is on drugs. Lots of drugs.



Taylor swift is apart of the Illuminati #grammys



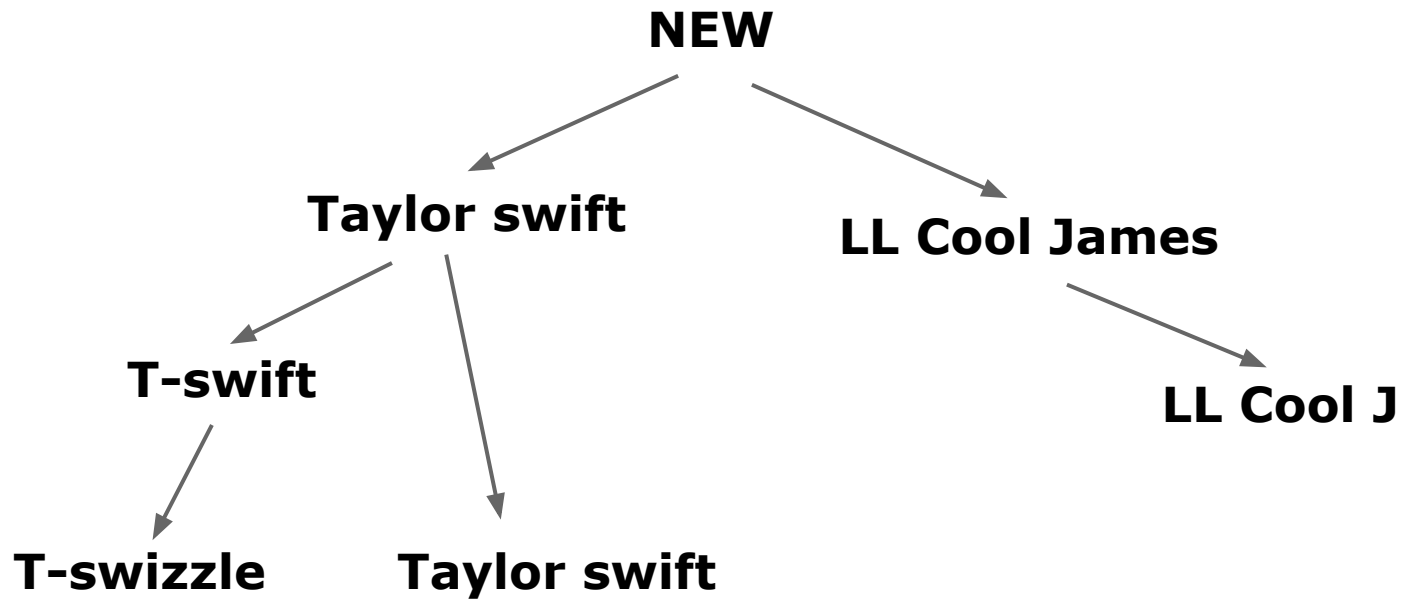
Ladies STILL love **LL Cool James**.



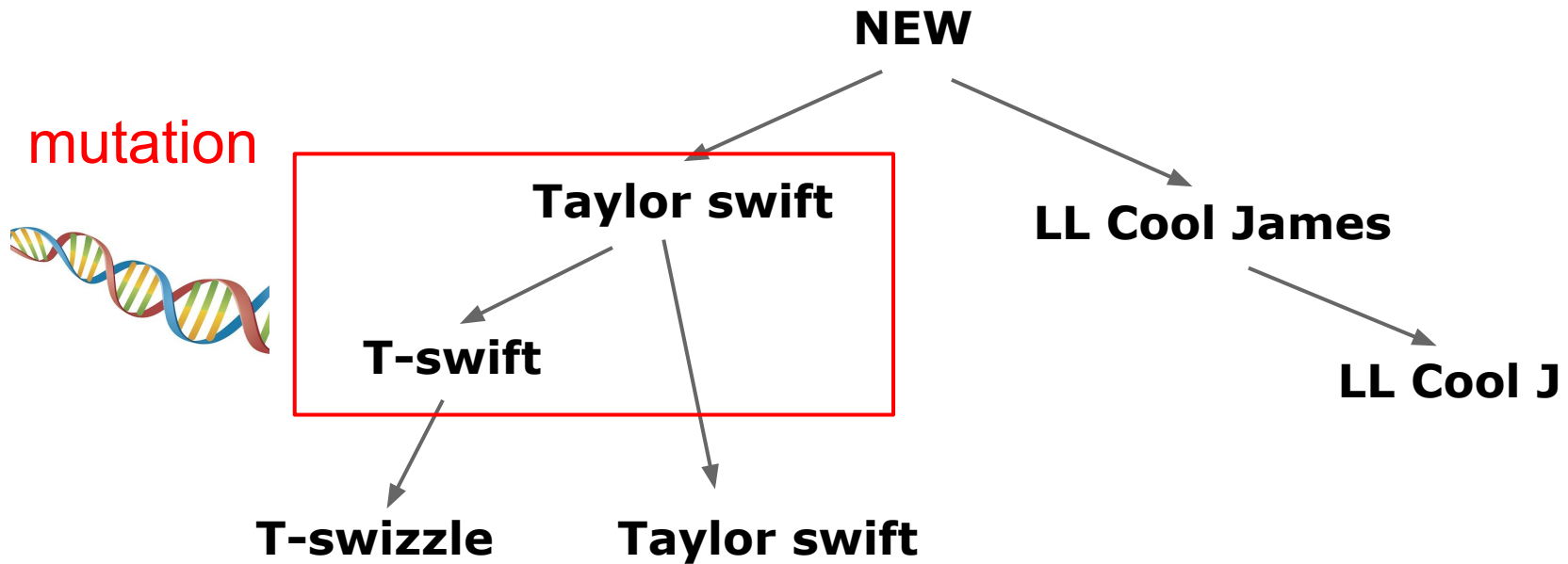
LL Cool J is looking reaIIlll chizzled!

Abbreviation

“phylogeny”?



“phylogeny”?



“phylogeny”?

species



NEW

Taylor swift

T-swift

T-swizzle

Taylor swift

LL Cool James

LL Cool J

A Generative Story



Did [**Taylor swift**] just dis harry sytles



Lets see how bad [**T Swift**] will be. #grammys



it's clear that [**T-Swizzle**] is on drugs



[**Taylor swift**] is apart of the Illuminati



Ladies STILL love [**LL Cool James**].



[**LL Cool J**] is looking realllll chizzled!

A Generative Story



Did **Taylor swift** just dis harry sytles



Lets see how bad **T Swift** will be. #grammys



it's clear that **T-Swizzle** is on drugs



Taylor swift is apart of the Illuminati



Ladies STILL love **LL Cool James** .



LL Cool J is looking reaIIlll chizzled!

Pick topics



1 5 2 1 2 3

Did **Taylor swift** just dis harry sytles



3 4 1 2 5 1 2 7

Lets see how bad **T Swift** will be. #grammys



3 1 2 5 1 2 4

it's clear that **T-Swizzle** is on drugs



5 1 3 2 3 8

Taylor swift is apart of the Illuminati



1 2 3 10

Ladies STILL love **LL Cool James** .



10 2 5 1 7

LL Cool J is looking realllll chizzled!

[] is a placeholder for an entity mention

Pick words | topics



1 5 2 1 2 3

Did [Taylor swift] just dis harry sytles



3 4 1 2 5 1 2 7

Lets see how bad [T Swift] will be. #grammys



3 1 2 5 1 2 4

it's clear that [T-Swizzle] is on drugs



5 1 3 2 3 8

[Taylor swift] is apart of the Illuminati



1 2 3 10

Ladies STILL love [LL Cool James].



10 2 5 1 7

[LL Cool J] is looking realllll chizzled!

Talk about NEW entity or an old one?

NEW

5



Did [**Taylor?** swift] just dis harry sytles

5



Lets see how bad [**T Swift**] will be. #grammys

5



it's clear that [**T-Swizzle**] is on drugs

5



[**Taylor swift**] is apart of the Illuminati

10



Ladies STILL love [**LL Cool James**].

10



[**LL Cool J**] is looking realllll chizzled!

Talk about NEW entity or an old one?

5

NEW



Did [**Taylor swift**] just dis harry sytles

5



Lets see how bad [**T Swift**] will be. #grammys

5



it's clear that [**T-Swizzle**] is on drugs

5



[**Taylor swift**] is apart of the Illuminati

10



Ladies STILL love [**LL Cool James**].

10



[**LL Cool J**] is looking realllll chizzled!

Talk about NEW entity or an old one?



5

Did [**Taylor swift**] just dis harry sytles

NEW



5

Lets see how bad [**T Swift ?**] will be. #grammys



5

it's clear that [**T-Swizzle**] is on drugs



5

[**Taylor swift**] is apart of the Illuminati



10

Ladies STILL love [**LL Cool James**].



10

[**LL Cool J**] is looking realllll chizzled!

Talk about NEW entity or an old one?



5

Did [**Taylor swift**] just dis harry sytles

NEW



5

Lets see how bad [**T Swift**] will be. #grammys



5

it's clear that [**T-Swizzle**] is on drugs



5

[**Taylor swift**] is apart of the Illuminati



10

Ladies STILL love [**LL Cool James**].



10

[**LL Cool J**] is looking realllll chizzled!

Talk about NEW entity or an old one?



5

Did [**Taylor swift**] just dis harry sytles

NEW

5

Lets see how bad [**T Swift**] will be. #grammys

5

it's clear that [**T-Swizzle**] is on drugs

5

[**Taylor swift**] is apart of the Illuminati

10

Ladies STILL love [**LL Cool James**].

10

[**LL Cool J**] is looking realllll chizzled!

Talk about NEW entity or an old one?



5

Did [**Taylor swift**] just dis harry sytles

NEW



5

Lets see how bad [**T Swift**] will be. #grammys



5

it's clear that [**T-Swizzle**] is on drugs



5

[**Taylor swift**] is apart of the Illuminati



10

Ladies STILL love [**LL Cool James**].



10

[**LL Cool J**] is looking realllll chizzled!

Talk about NEW entity or an old one?



5
Did [**Taylor swift**] just dis harry sytles

5
Lets see how bad [**T Swift**] will be. #grammys

5
it's clear that [**T-Swizzle**] is on drugs

5
[**Taylor?** swift] is apart of the Illuminati

10
Ladies STILL love [**LL Cool James**].

10
[**LL Cool J**] is looking realllll chizzled!

NEW

Talk about NEW entity or an old one?



5

Did [**Taylor swift**] just dis harry sytles

NEW

5

Lets see how bad [**T Swift**] will be. #grammys

5

it's clear that [**T-Swizzle**] is on drugs

5

[**Taylor swift**] is apart of the Illuminati

10

Ladies STILL love [**LL Cool James**].

10

[**LL Cool J**] is looking realllll chizzled!

Selecting a parent

$$p(\text{parent}) = \frac{\exp(\mathbf{f}(\text{parent}, \text{child}) \cdot \phi)}{Z}$$



5

[Taylor swift] is apart of the Illuminati

NEW



10

Ladies STILL love [LL Cool James].



10

[LL Cool J] is looking realllll chizzled!

Selecting a parent

Unknown parameters

$$p(\text{parent}) = \frac{\exp(\phi_{10,10} + \phi_{\text{equal}})}{Z}$$



5

[Taylor swift] is apart of the Illuminati

NEW



10

Ladies STILL love [LL Cool James].



10

[LL Cool J] is looking realllll chizzled!

Selecting a parent

$$p(\text{parent}) = \frac{\exp(\phi_{5,10} + \phi_{\text{diff}})}{Z}$$



5

[Taylor swift] is apart of the Illuminati



10

Ladies STILL love [LL Cool James].



10

[LL Cool J] is looking realllll chizzled!

NEW

Selecting a parent

Hyperparameter:
controls # of entities

$$p(\text{parent}) = \frac{\exp(\phi_{\text{new}})}{Z}$$



5

[Taylor swift] is apart of the Illuminati



Ladies STILL love [LL Cool James].

10



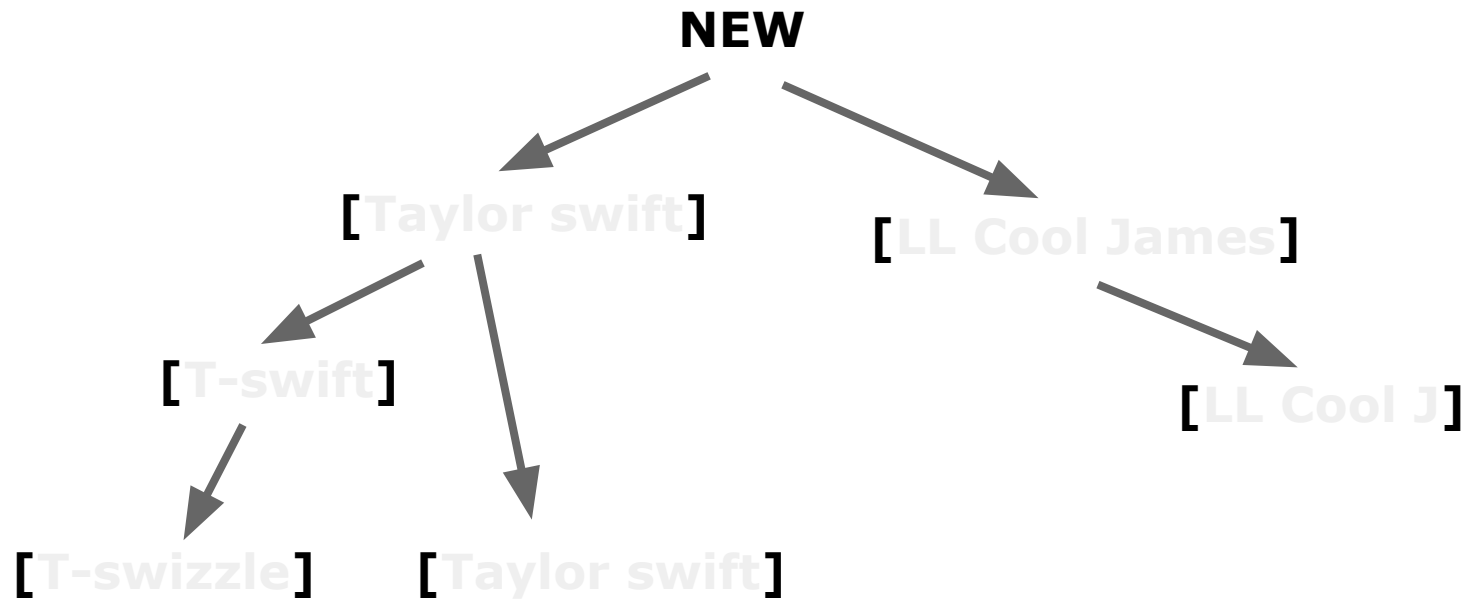
10

[LL Cool J] is looking realllll chizzled!

NEW

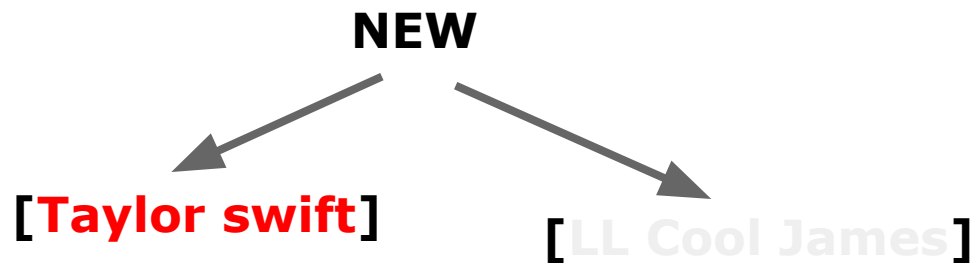
Another view

Note: No names yet, that's next...



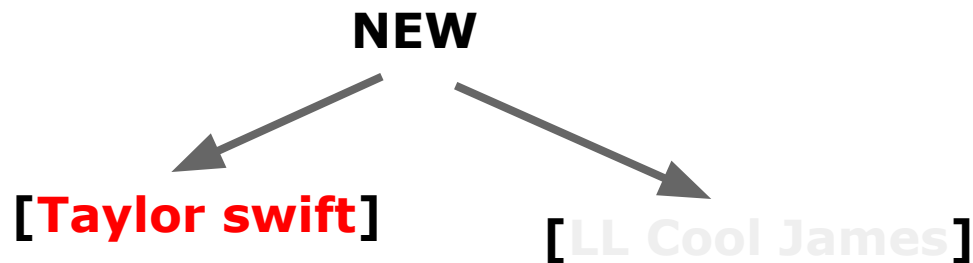
Generating a name

If parent = NEW: name new entity



Generating a name

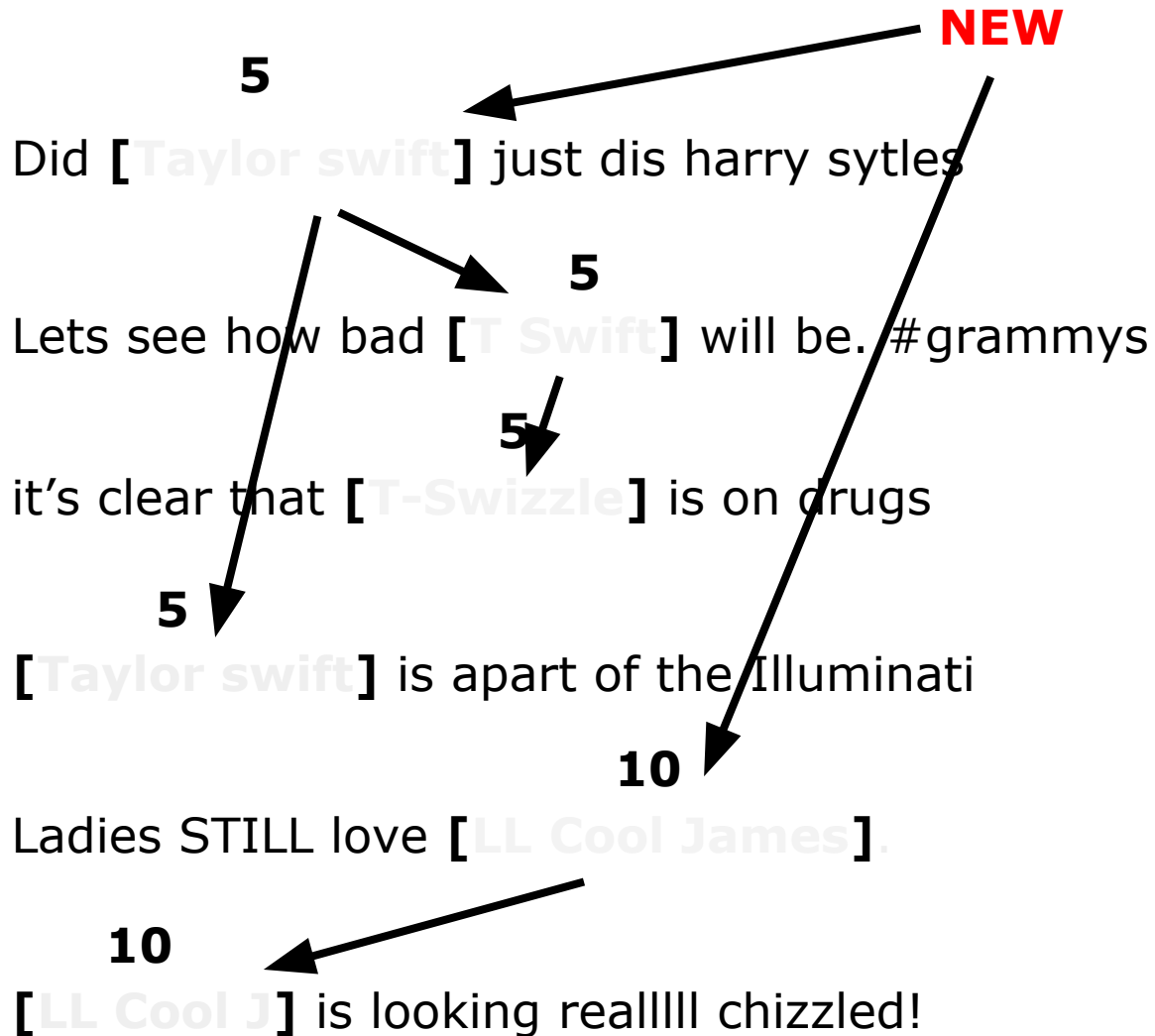
If parent = NEW: name new entity



We draw a name from a simple character LM with trainable parameters θ

(... room for fancier distributions over Σ^* ...)

Pick mention name | parent



Pick mention name | parent



Did [**Taylor swift**] just dis harry sytles

Lets see how bad [**T Swift**] will be. #grammys

it's clear that [**T-Swizzle**] is on drugs

[**Taylor swift**] is apart of the Illuminati

Ladies STILL love [**LL Cool James**].

[**LL Cool J**] is looking realllll chizzled!

NEW

Pick mention name | parent



Did [**Taylor swift**] just dis harry sytles

Lets see how bad [**T Swift**] will be. #grammys

it's clear that [**T-Swizzle**] is on drugs

[**Taylor swift**] is apart of the Illuminati


Ladies STILL love [**LL Cool James**]

[**LL Cool J**] is looking realllll chizzled!

NEW

Generating a name

If parent \neq NEW: copy (maybe mutate) the parent's name

[Taylor swift]

[T-Swift]

We draw the name from a stochastic contextual edit model (see Cotterell et al., 2014) with trainable parameters θ

Taylor swift

Tεεεεε-Swift


Copy Del

Subst Copy

Stop

Generating a name

If parent \neq NEW: copy (maybe mutate) the parent's name

[Taylor swift]

[**T-Swift**]

We draw the name from a stochastic contextual edit model (see Cotterell et al., 2014) with trainable parameters θ

θ gives the contextual probability of different character edits

- copy, delete, substitute c , insert c

Pick mention name | parent



Did [**Taylor swift**] just dis harry sytles

Lets see how bad [**T Swift**] will be. #grammys

it's clear that [**T-Swizzle**] is on drugs

[**Taylor swift**] is apart of the Illuminati

Ladies STILL love [**LL Cool James**]

[**LL Cool J**] is looking realllll chizzled!

NEW

Pick mention name | parent



Did [**Taylor swift**] just dis harry sytles

Lets see how bad [**T Swift**] will be. #grammys

it's clear that [**T-Swizzle**] is on drugs

[**Taylor swift**] is apart of the Illuminati

Ladies STILL love [**LL Cool James**]

[**LL Cool J**] is looking realllll chizzled!

NEW

Pick mention name | parent



Did [**Taylor swift**] just dis harry sytles

Lets see how bad [**T Swift**] will be. #grammys

it's clear that [**T-Swizzle**] is on drugs

[**Taylor swift**] is apart of the Illuminati

Ladies STILL love [**LL Cool James**]

[**LL Cool J**] is looking realllll chizzled!

NEW

Pick mention name | parent



Did [**Taylor swift**] just dis harry sytles

Lets see how bad [**T Swift**] will be. #grammys

it's clear that [**T-Swizzle**] is on drugs

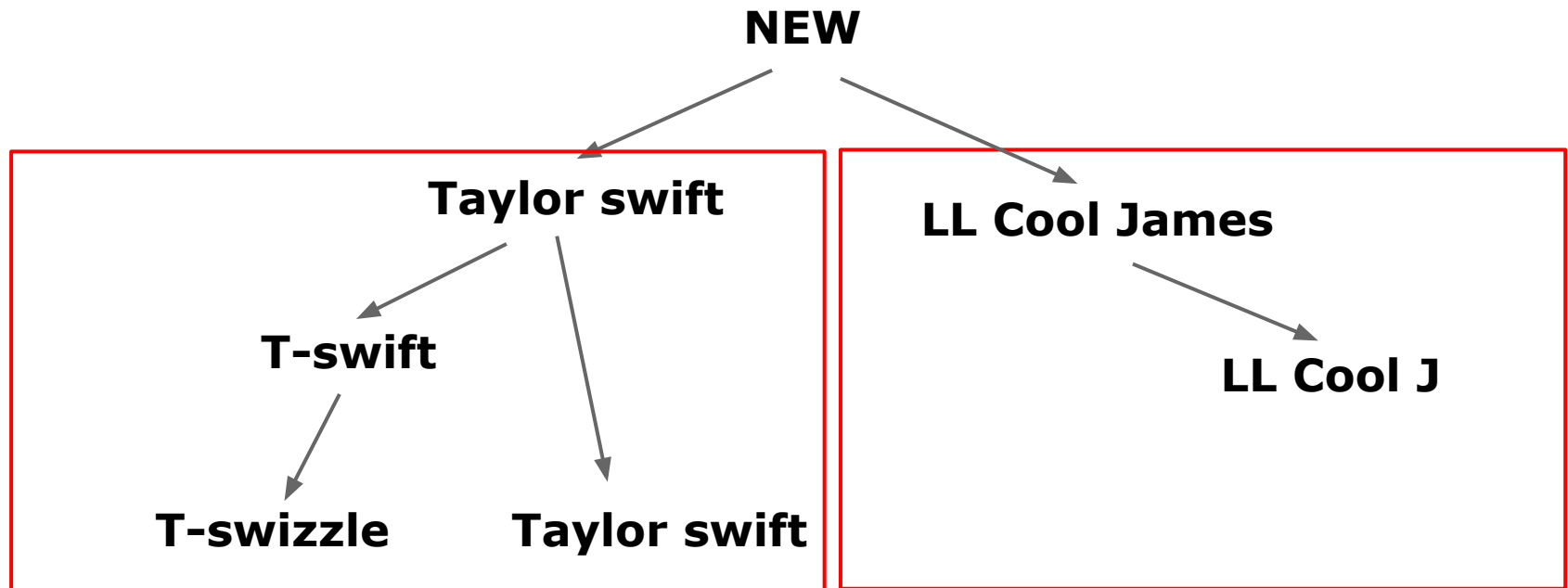
[**Taylor swift**] is apart of the Illuminati

Ladies STILL love [**LL Cool James**]

[**LL Cool J**] is looking realllll chizzled!

NEW

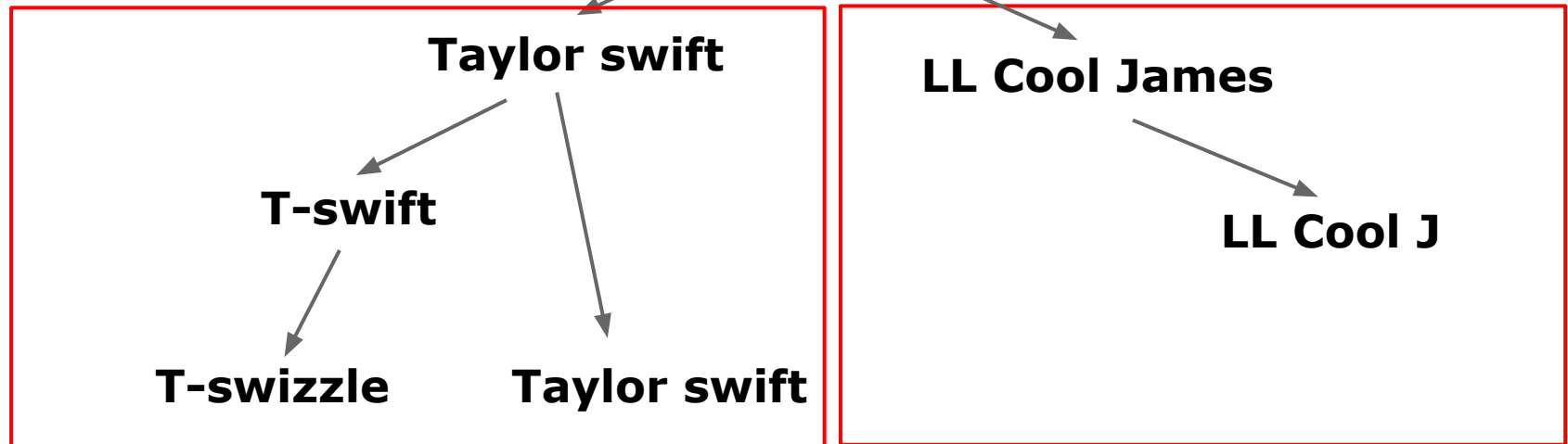
End result



End result



NEW



Inference (fixed parameters)



Did [**Taylor swift**] just dis harry sytles



Lets see how bad [**T Swift**] will be. #grammys



it's clear that [**T-Swizzle**] is on drugs



[**Taylor swift**] is apart of the Illuminati



Ladies STILL love [**LL Cool James**].



[**LL Cool J**] is looking realllll chizzled!

Many possible phylogenies



Did **[Taylor swift]** just dis harry sytles

Lets see how bad **[T Swift]** will be. #grammys

it's clear that **[T-Swizzle]** is on drugs

[Taylor swift] is apart of the Illuminati

Ladies STILL love **[LL Cool James]**.

[LL Cool J] is looking realllll chizzled!

NEW



Many possible phylogenies



Did **[Taylor swift]** just dis harry sytles

Lets see how bad **[T Swift]** will be. #grammys

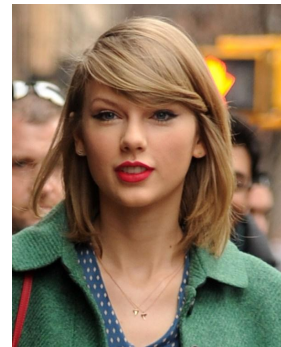
it's clear that **[T-Swizzle]** is on drugs

[Taylor swift] is apart of the Illuminati

Ladies STILL love **[LL Cool James]**.

[LL Cool J] is looking realllll chizzled!

NEW



Many possible phylogenies



Did **[Taylor swift]** just dis harry sytles

Lets see how bad **[T Swift]** will be. #grammys

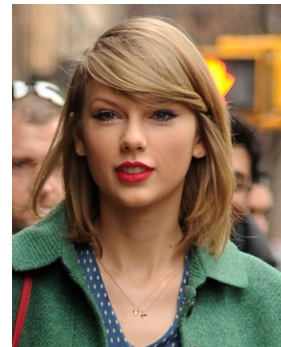
it's clear that **[T-Swizzle]** is on drugs

[Taylor swift] is apart of the Illuminati

Ladies STILL love **[LL Cool James]**.

[LL Cool J] is looking realllll chizzled!

NEW



Many possible phylogenies



Did **[Taylor swift]** just dis harry sytles

Lets see how bad **[T Swift]** will be. #grammys

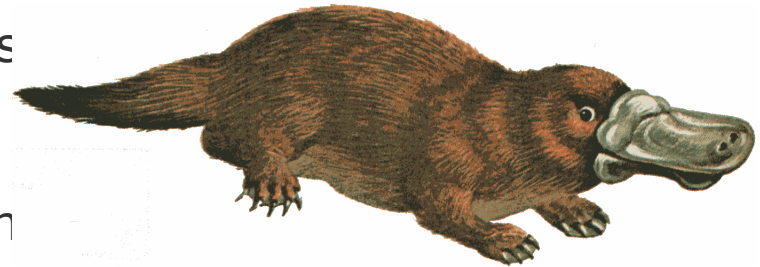
it's clear that **[T-Swizzle]** is

[Taylor swift] is apart of th

Ladies STILL love **[LL Cool James]**.

[LL Cool J] is looking realllll chizzled!

NEW



Many possible phylogenies



Did **[Taylor swift]** just dis harry sytles

Lets see how bad **[T Swift]** will be. #grammys

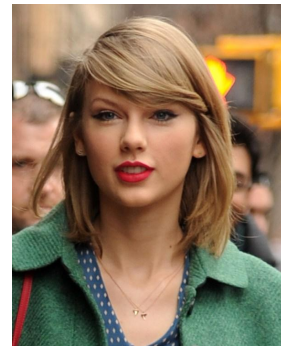
it's clear that **[T-Swizzle]** is on drugs

[Taylor swift] is apart of the Illuminati

Ladies STILL love **[LL Cool James]**.

[LL Cool J] is looking realllll chizzled!

NEW



Many possible topic assignments

5



Did [**Taylor swift**] just dis harry sytles

5



Lets see how bad [**T Swift**] will be. #grammys

5



it's clear that [**T-Swizzle**] is on drugs

5



[**Taylor swift**] is apart of the Illuminati

10



Ladies STILL love [**LL Cool James**].

10



[**LL Cool J**] is looking realllll chizzled!

Many possible topic assignments

1



Did [**Taylor swift**] just dis harry sytles

1



Lets see how bad [**T Swift**] will be. #grammys

2



it's clear that [**T-Swizzle**] is on drugs

3



[**Taylor swift**] is apart of the Illuminati

5



Ladies STILL love [**LL Cool James**].

2



[**LL Cool J**] is looking realllll chizzled!

Many possible topic assignments

3



Did [**Taylor swift**] just dis harry sytles

3



Lets see how bad [**T Swift**] will be. #grammys

3



it's clear that [**T-Swizzle**] is on drugs

3



[**Taylor swift**] is apart of the Illuminati

3



Ladies STILL love [**LL Cool James**].

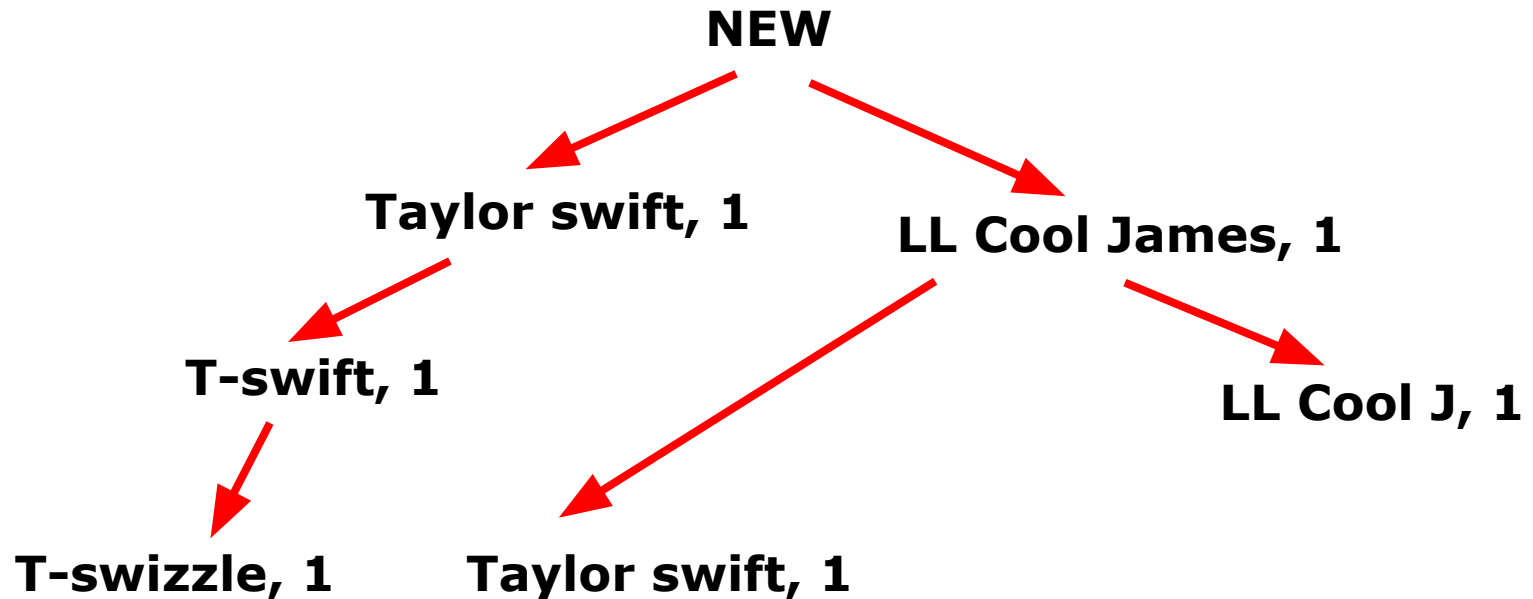
3



[**LL Cool J**] is looking realllll chizzled!

Block Gibbs (fixed parameters)

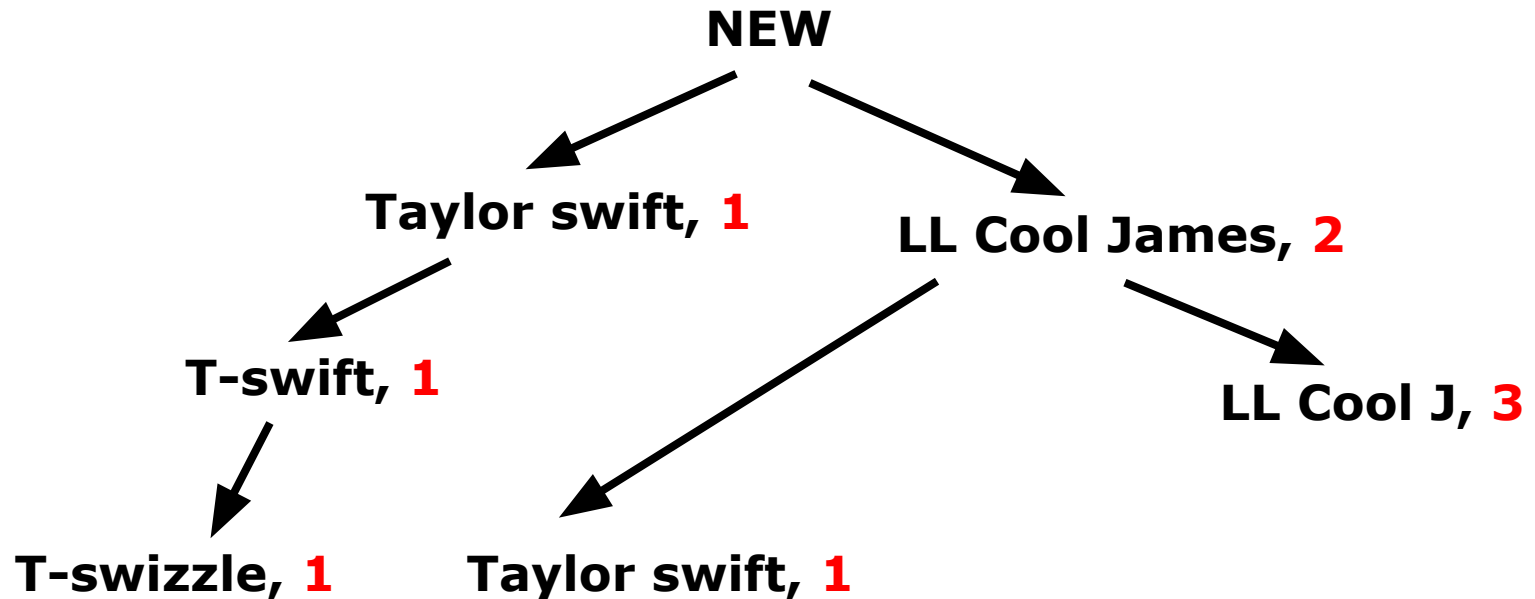
Sample **phylogeny** | topics



$$p(x \rightarrow y) \propto p_{\phi}(x \mid y) p_{\theta}(y.\text{name} \mid x.\text{name})$$

Block Gibbs (fixed parameters)

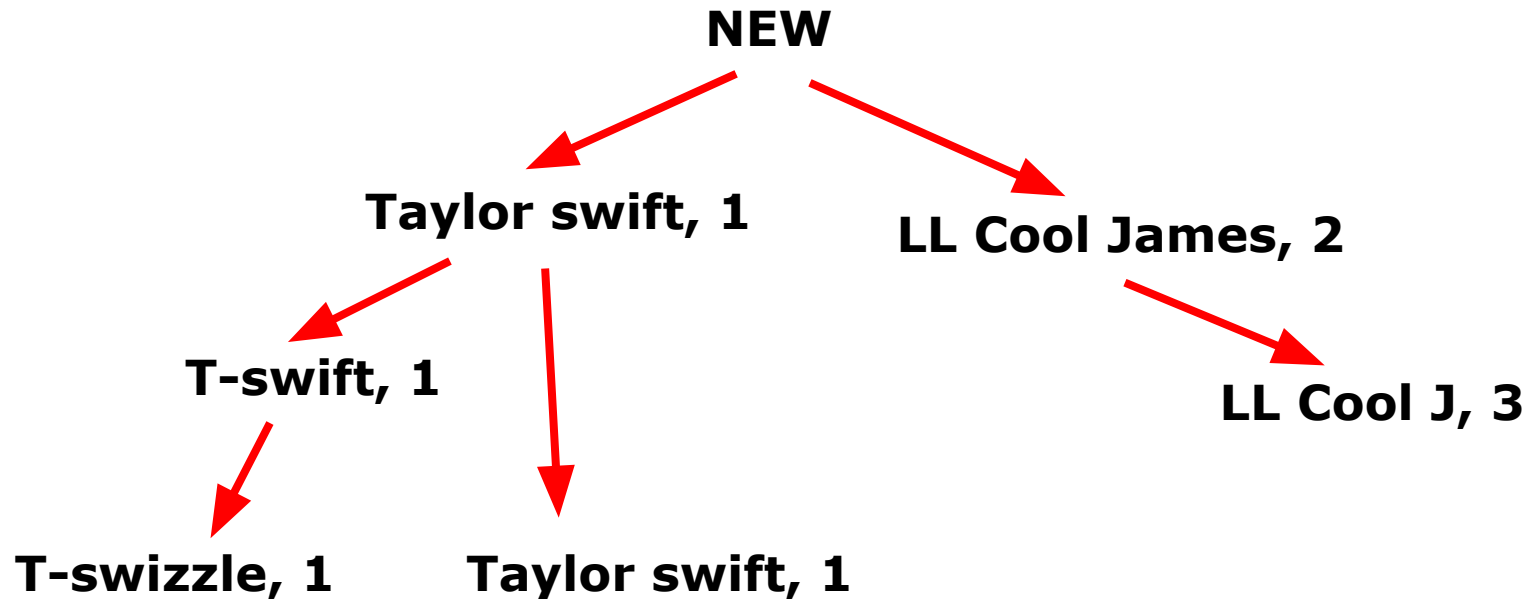
Sample **topics** | phylogeny



(We use tree-structured BP to construct proposals;
see paper for details.)

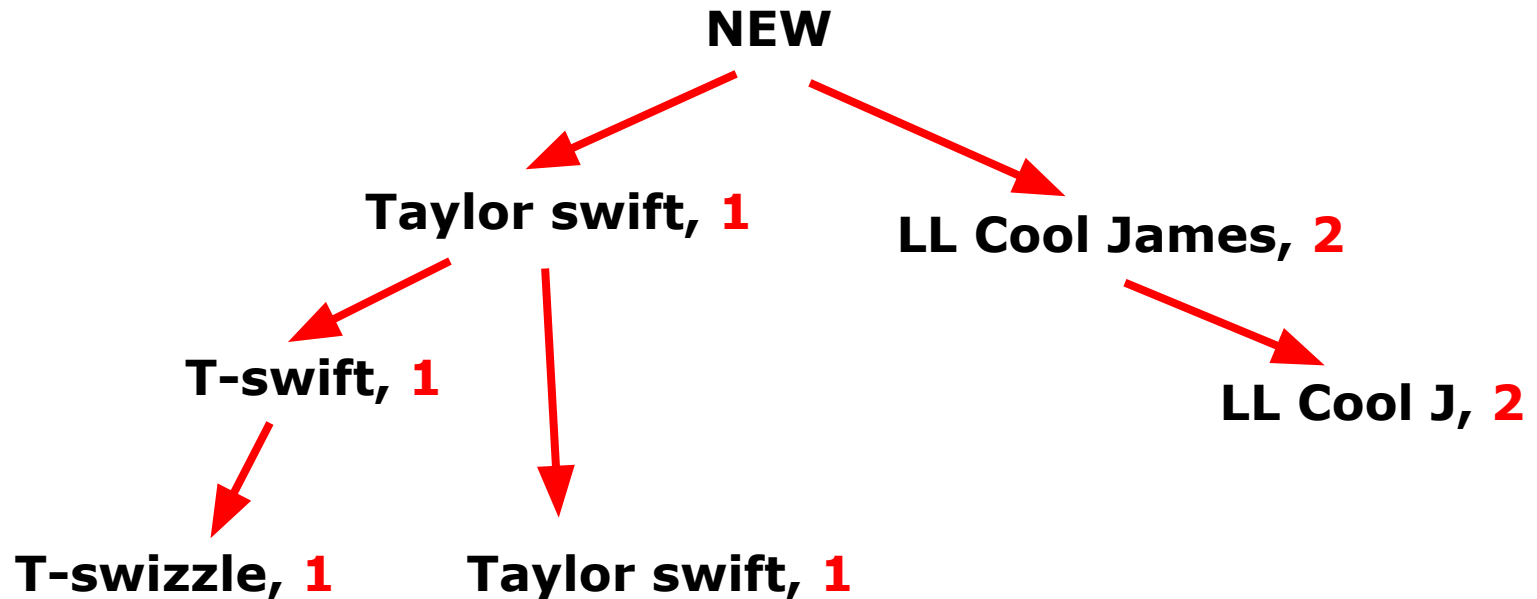
Block Gibbs (fixed parameters)

Sample **phylogeny** | topics

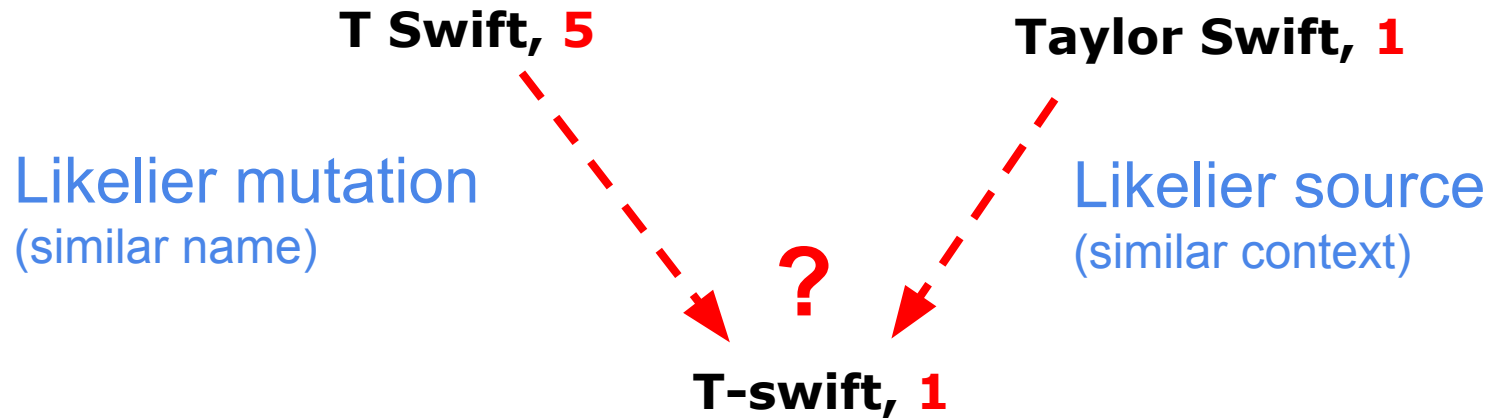


Block Gibbs (fixed parameters)

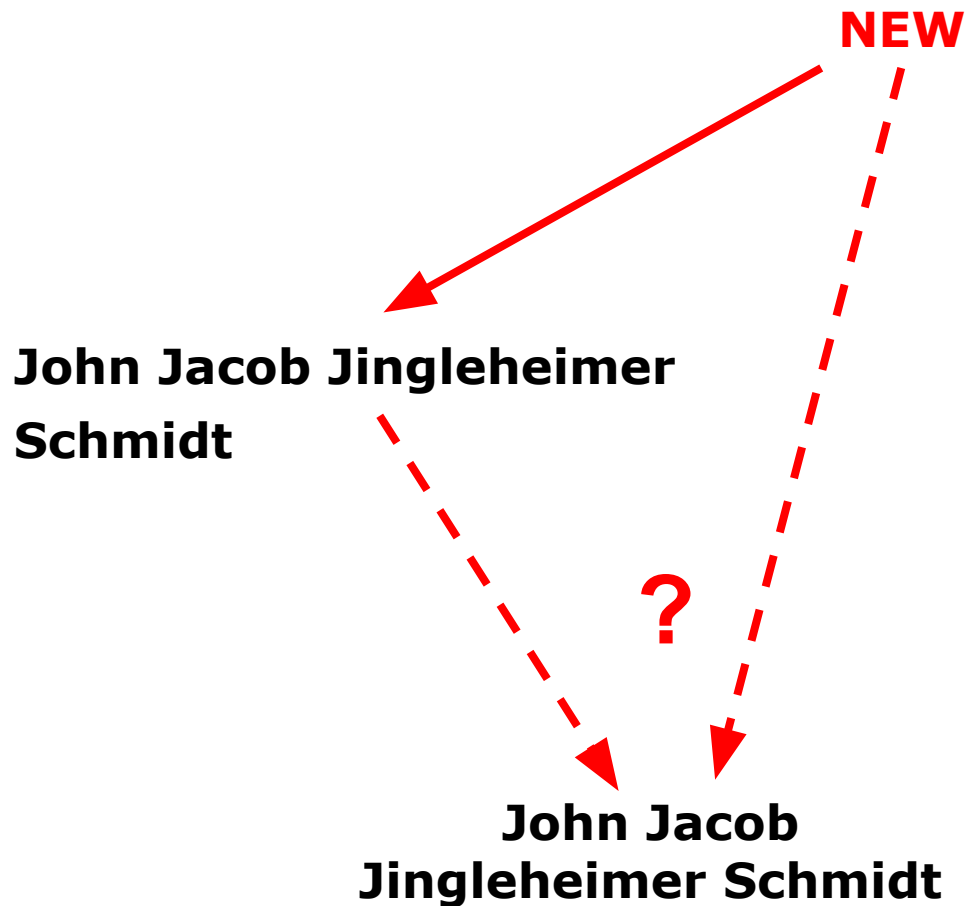
Sample **topics** | phylogeny



What is the sampler thinking about?



What is the sampler thinking about?



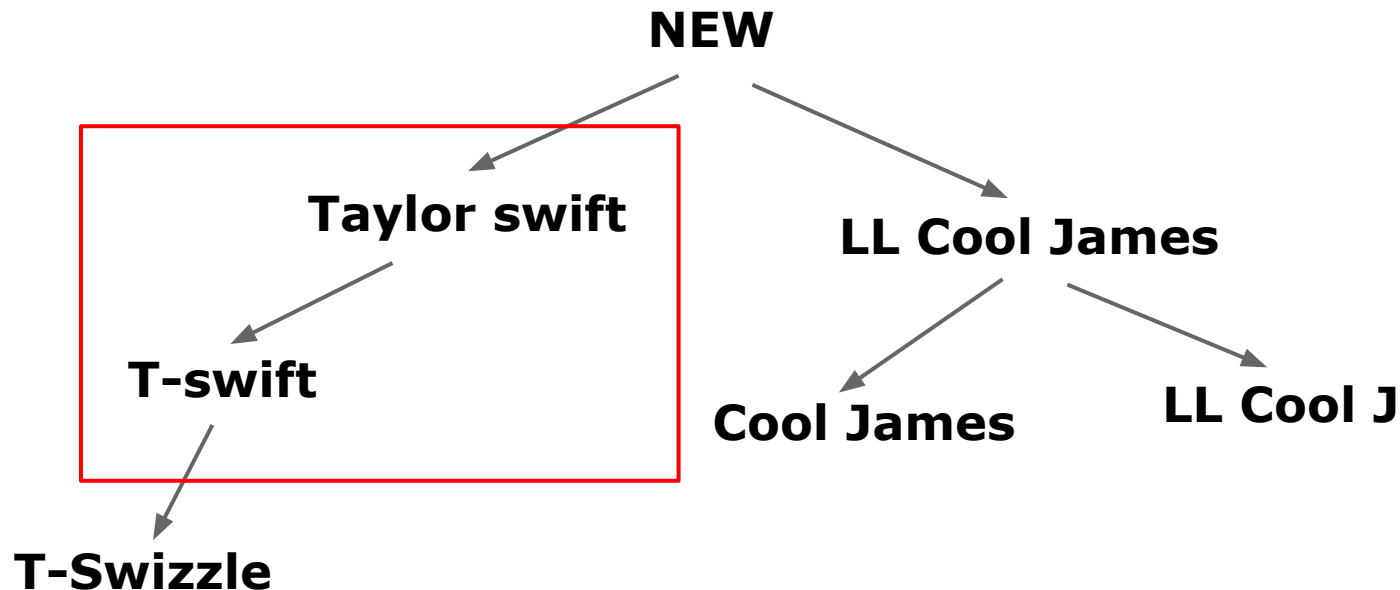
Prefer to **copy** rather than generate from scratch

(short or common names *would* be plausible to generate twice)

(see paper for an improved mutation model that considers pragmatics)

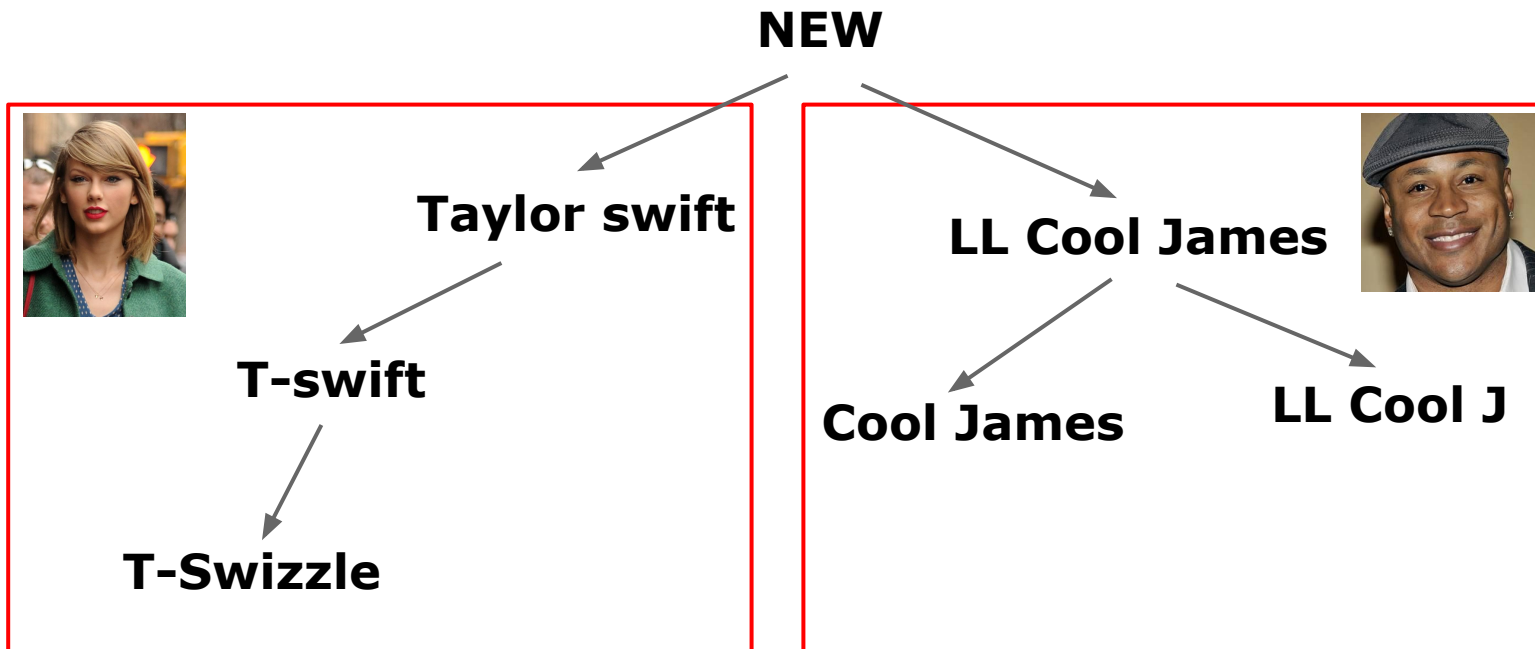
What do samples tell us?

1. Which names are copies of other names
(used for EM training of all parameters)



What do samples tell us?

2. Which names corefer (used for your IE task)



A complication

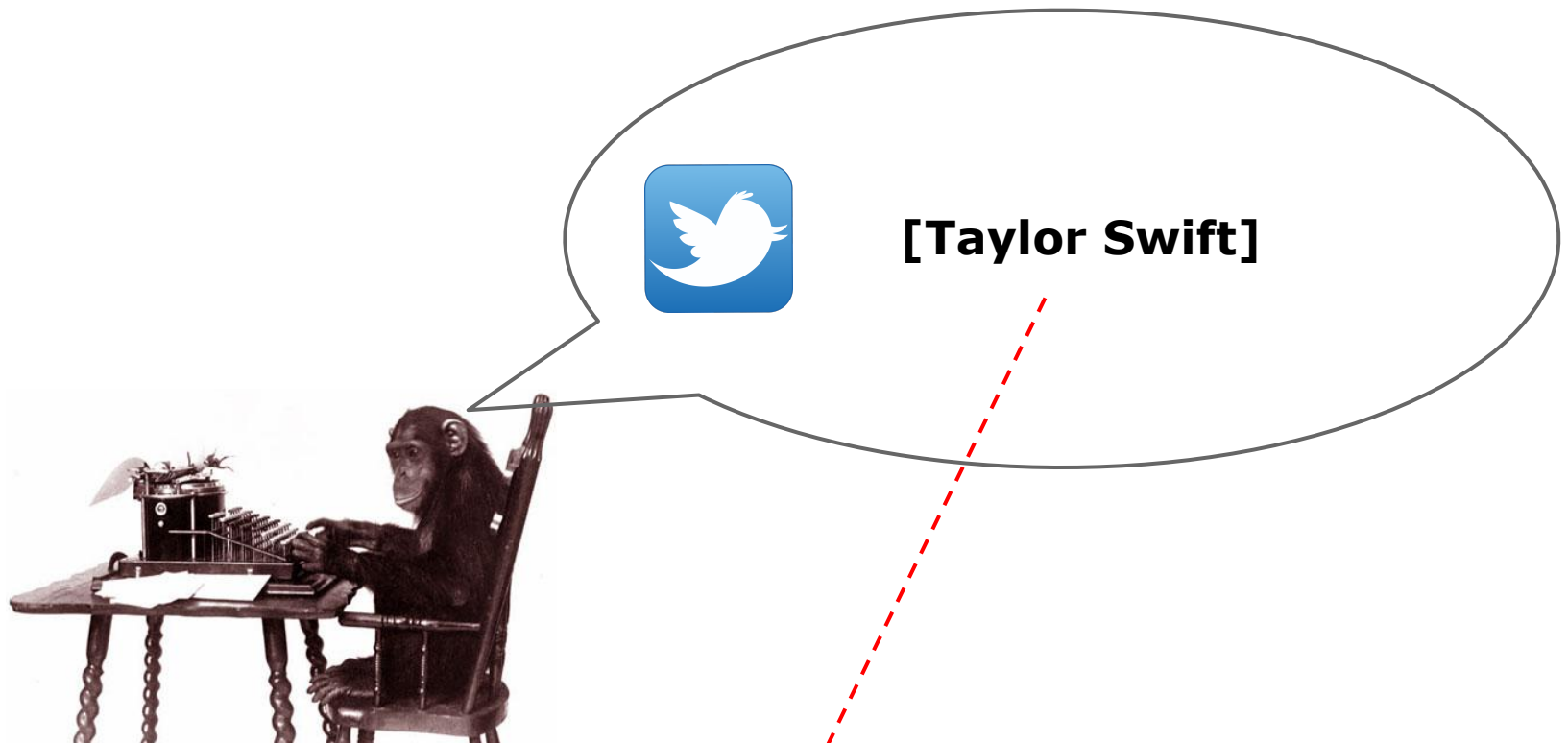


Wait a second...



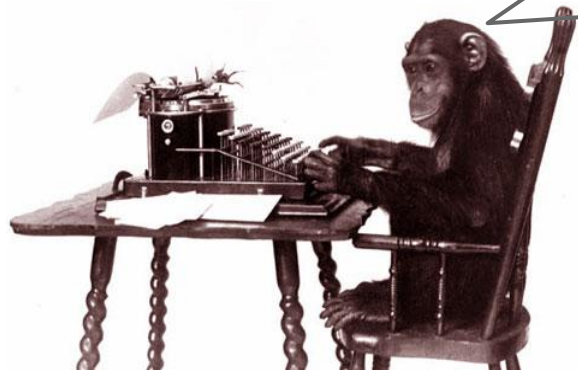
Lets see how bad **[T Swift]** will be. #grammys

Authors copy from multiple sources



Lets see how bad **[T Swift]** will be. #grammys

Authors copy from multiple sources

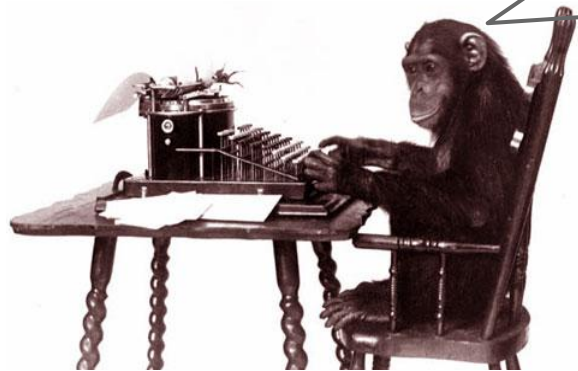


The
New York Times [Taylor Swift]
Times



Lets see how bad [T Swift] will be. #grammys

Authors copy from multiple sources



CNN

[Taylor Swift]



Lets see how bad **[T Swift]** will be. #grammys

Bad ordering



it's clear that **[T-Swizzle]** is on drugs

Did **[Taylor swift]** just dis harry sytles

Lets see how bad **[T Swift]** will be. #grammys

NEW

Bad ordering



[Taylor Swift]



[T Swift]

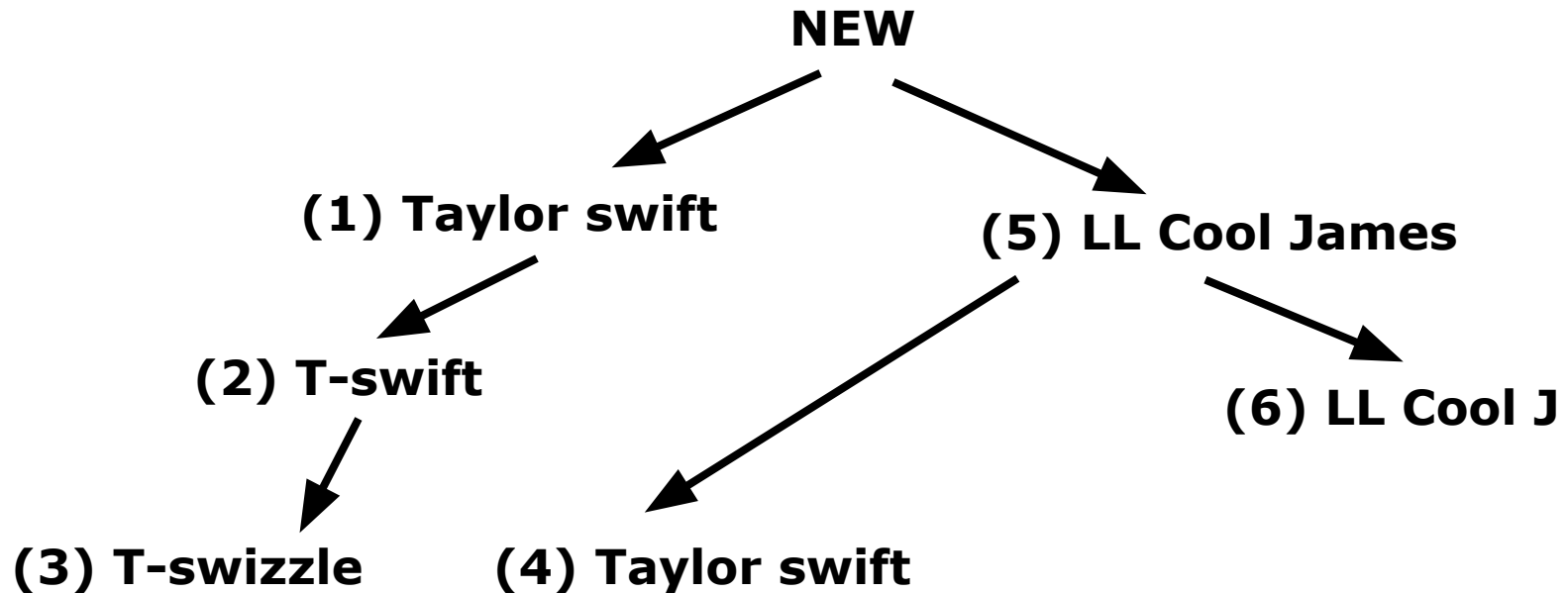


it's clear that [T-Swizzle] is on drugs

Did [Taylor swift] just dis harry sytles

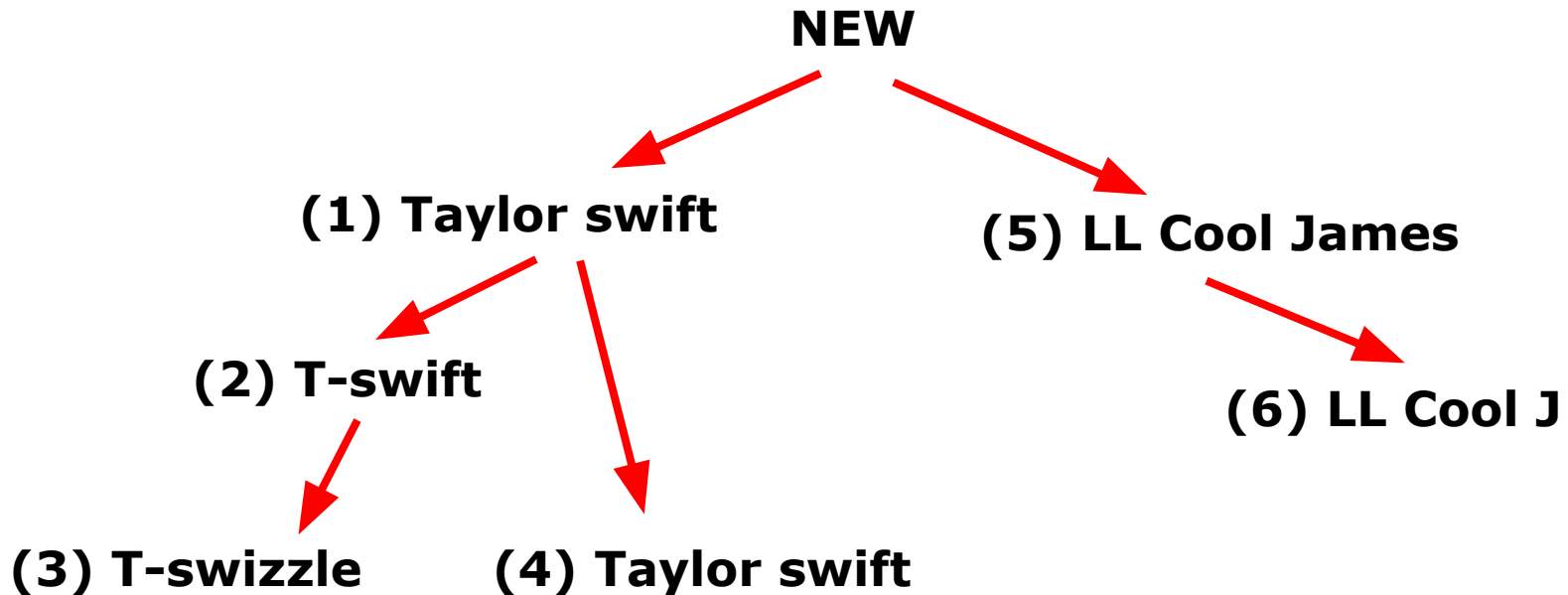
Lets see how bad [T Swift] will be. #grammys

Solution: Treat order as unknown



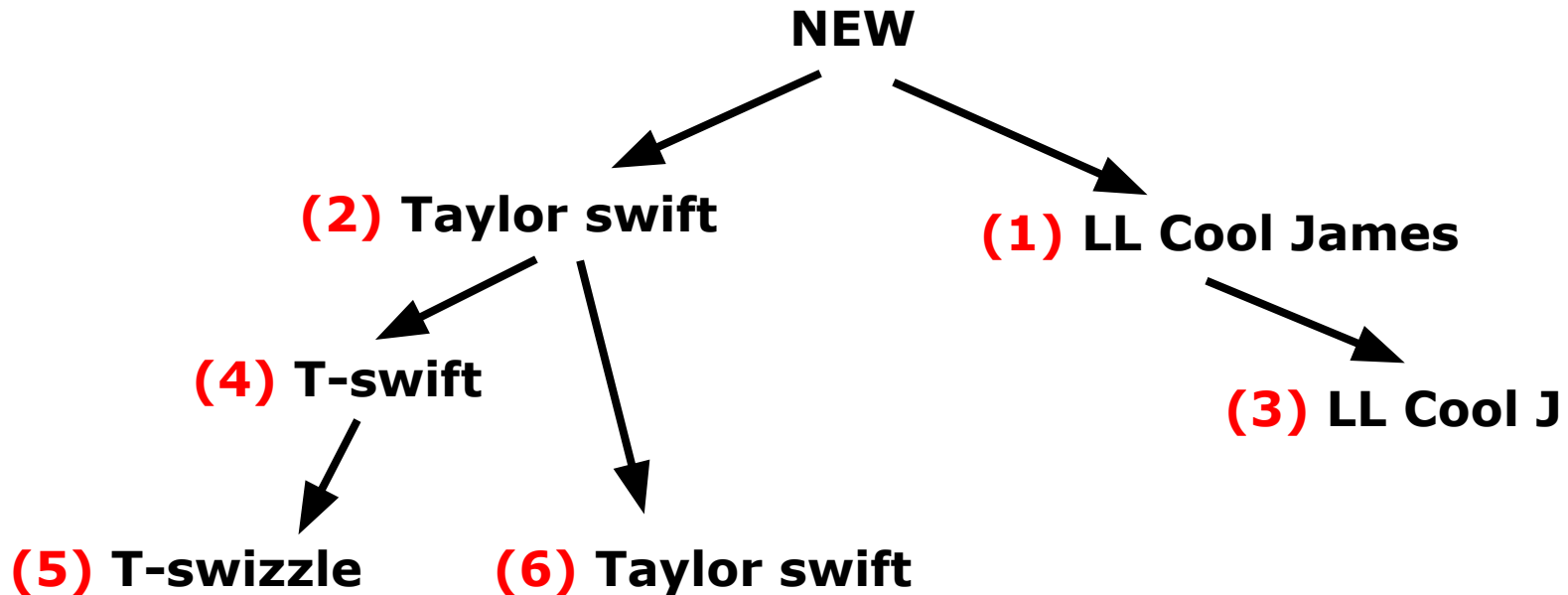
Updated Block Gibbs (fixed params)

Sample **phylogeny** | ordering, topics



Updated Block Gibbs (fixed params)

Sample **ordering** | phylogeny, topics



(we use a proposal distribution and correct with MH;
see paper for details)

Summary

So far we've seen:

- The generative story
- A sampler for posterior inference

Wrap up with:

- Parameter estimation
- MBR decoding
- Experiments

Parameter Estimation

Monte Carlo EM

Repeat:

- **E-Step:** Count edges in sample
- **M-Step:** Take stochastic gradient step
 - Update* parent model parameters
 - Update* mutation model parameters

* We use AdaGrad which has an adaptive learning rate

Consensus Clustering

To get a single “hard” clustering C for evaluation, we use minimum Bayes risk:

$$\operatorname{argmin}_C \mathbf{E}_{C'}[\operatorname{loss}(C, C')]$$

- Minimize expected loss of C
 - With respect to C' drawn from the model posterior
 - Estimate this using our samples of C'

(we approximate argmin using spectral clustering; see paper for details)

Summary

Unsupervised clustering procedure:

1. Train model using Monte Carlo EM
2. Sample from the posterior
3. Pick the MBR clustering given posterior

Evaluation: Twitter



Corpus of Twitter messages discussing the 2013 Grammy award ceremony (~5000 posts total, ~300 entities).

- Procedure: 4-fold cross validation
 - **Train:** Tune weight of picking NEW as the parent to control precision / recall trade-off
 - **Test:** Run clustering procedure with this weight fixed

Results: Twitter

System	B³ F1
Exact-Match	69.8

(averages over 4 folds)

Results: Twitter

System	B³ F1
Exact-Match	69.8
Green et al. (2012)	79.3

(averages over 4 folds)

Results: Twitter

System	B³ F1
Exact-Match	69.8
Green et al. (2012)	79.3
Phylo (no context)	88.7

(averages over 4 folds)

Results: Twitter

System	B³ F1
Exact-Match	69.8
Green et al. (2012)	79.3
Phylo (no context)	88.7
Phylo + Topic	91.8

(averages over 4 folds)

Results: Twitter

System	B³ F1
Exact-Match	69.8
Green et al. (2012)	79.3
Phylo (no context)	88.7
Phylo + Topic	91.8
Phylo + Topic + MBR	91.9

(averages over 4 folds)

Results: ACE 2008 **AP**



Corpus of news articles, mostly politics
~4000 mentions, ~2000 entities

- Procedure

Train: Tune weight of picking NEW as the parent to control precision / recall trade-off

Test: Run clustering procedure with this weight fixed

Evaluation: ACE 2008

System	PER B3 F1	ORG B3 F1
Exact-Match	88.8	87.1

Evaluation: ACE 2008

System	PER B3 F1	ORG B3 F1
Exact-Match	88.8	87.1
Green et al. (2012)	91.9	90.3

Evaluation: ACE 2008

System	PER B3 F1	ORG B3 F1
Exact-Match	88.8	87.1
Green et al. (2012)	91.9	90.3
Phylo+Topic+MBR	92.7	87.6

Thanks!

- More experiments in the paper
 - Name canonicalization
- Code will be released soon
 - <https://bitbucket.org/noandrews/phyloinf>
- Future uses of model and code?
 - Track diffusion of memes through social media
 - Derivational morphology

Samples tell us which entities corefer



Useta love reading old **[Tom Swift]** books



Did **[Taylor swift]** just dis harry sytles



Lets see how bad **[T Swift]** will be. #grammys



it's clear that **[T-Swizzle]** is on drugs

More about
this person?



[Taylor swift] is apart of the Illuminati



Ladies STILL love **[LL Cool James]**.

Samples tell us which entities corefer



Useta love reading old **[Tom Swift]** books



Did **[Taylor swift]** just dis harry sytles



Lets see how bad **[T Swift]** will be. #grammys



it's clear that **[T-Swizzle]** is on drugs

More about
this person?

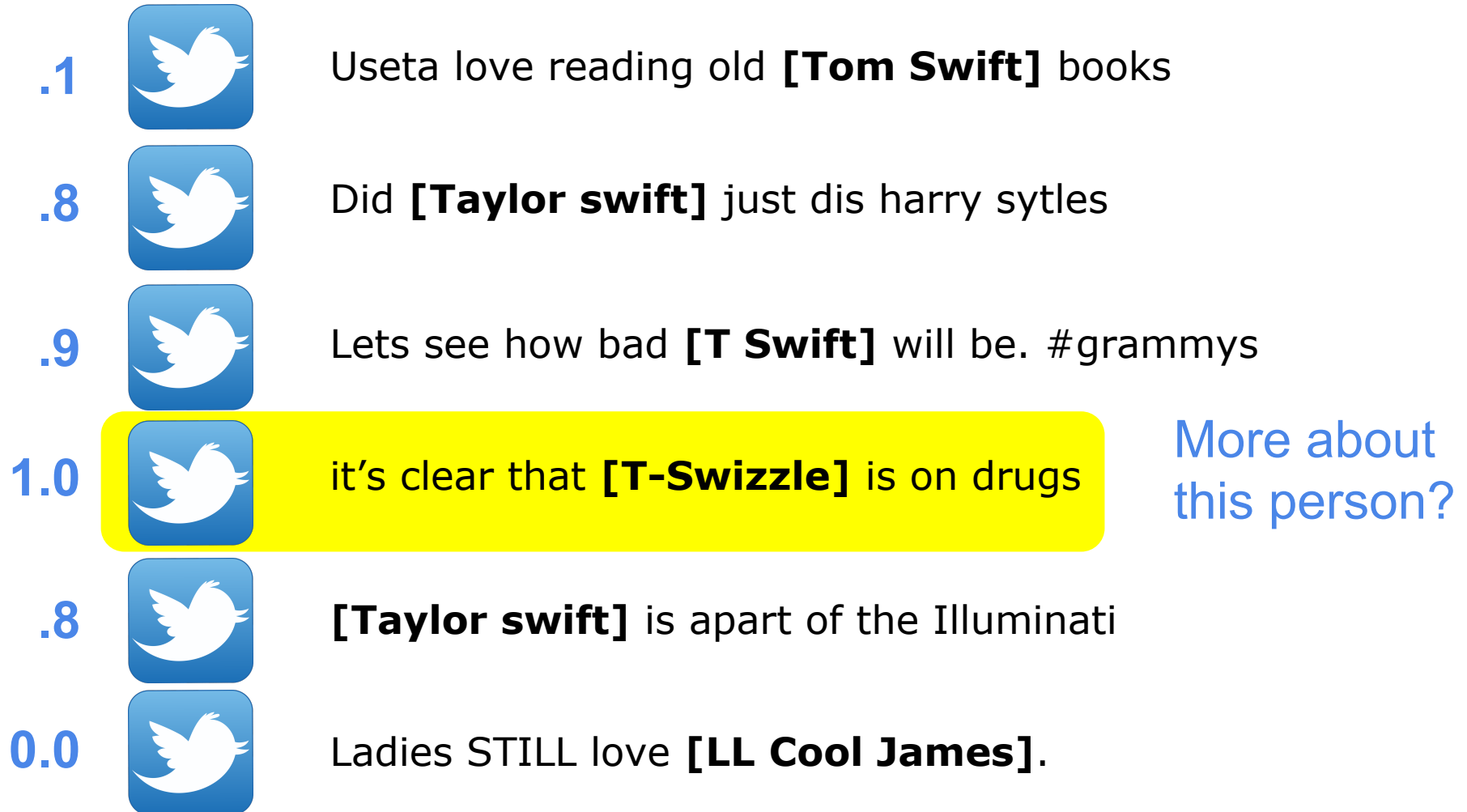


[Taylor swift] is apart of the Illuminati



Ladies STILL love **[LL Cool James]**.

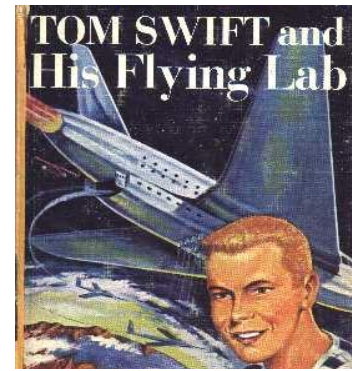
Coref probabilities from many samples



.8



.1



.1



Useta love reading old **[Tom Swift]** books

.8



Did **[Taylor swift]** just dis harry sytles

.9



Lets see how bad **[T Swift]** will be. #grammys

1.0



it's clear that **[T-Swizzle]** is on drugs

More about
this person?

.8



[Taylor swift] is apart of the Illuminati

0.0



Ladies STILL love **[LL Cool James]**.