Microblog Summarization Using Conversation Structures

Presenter: Jing Li

Nov 26, 2016

Outline

- Background
- Microblog Topic Extraction
- Conversation Tree Summarization
- Conclusion

Outline

- **Background**
- Microblog Topic Extraction
- Conversation Tree Summarization
- Conclusion

Background

- Microblog: center for reporting, discussing and disseminating real-life issues.
    - e.g., UEFA European Championship, terrorist attacks in Paris, etc.
- Millions of messages streaming out everyday.
- Microblog summarization
    - Microblog posts are short and informal rendering the lack of context information

Background

- Microblog: center for reporting, discussing and disseminating real-life issues.
- Millions of messages streaming out everyday.
- Microblog summarization
    - Microblog posts are short and informal rendering the lack of context information
    - Conversation trees based on reposting/replying relations

Background

- Given a collection with multiple conversation trees covering various topics.
- How do we effectively summarize?
- Cluster messages into different topics
    - Topic models, e.g., LDA

Outline

- Background
- **Microblog Topic Extraction**
- Conversation Tree Summarization
- Conclusion

Outline

- **Microblog Topic Extraction**
  - Introduction
  - Conversation Modeling
  - LeadLDA Topic Model
  - Experiments

Outline

- **Microblog Topic Extraction**
  - **Introduction**
  - Conversation Modeling
  - LeadLDA Topic Model
  - Experiments

Topic Models on Microblog

- Why do we extract topics from microblog?
  - Extract topics represented as word distributions
  - Uncover the hidden semantic structures
  - Useful to downstream applications, e.g., summarization
- Is it a challenging problem?
  - Microblog posts are short and colloquial

Topic Models on Microblog

- Why do we extract topics from microblog?
- Is it a challenging problem?
    - Microblog posts are short and colloquial
        - Sparsity of document-level word co-occurrence
        - Non-topic words are common

Prior Works

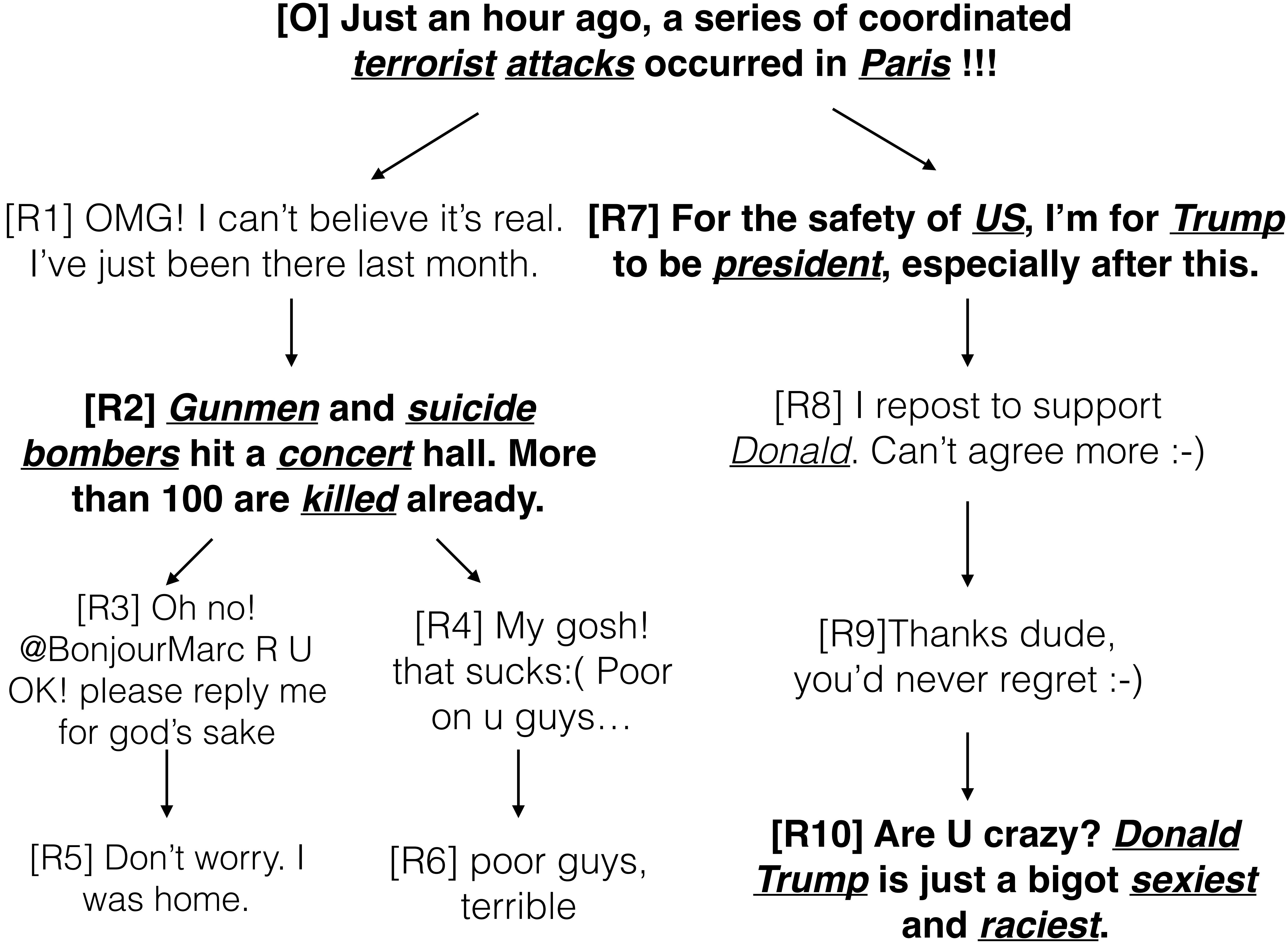
- Aggregate messages.
    - Authorship, shared words, hashtags, etc.
    - These strategies are suboptimal
- Directly model topics of biterms in posts (Yan et al. 13)
    - The context is still “short”
- Aggregate texts jointly with topic inference (Quan et al. 15)
    - No prior information given to text aggregation

Our Idea

- Conversations on microblog
  - Reposts and replies
- Organize posts as conversation trees.
  - Enrich contextual information
  - Provide clues to identify key words for topic representation

Outline

- **Microblog Topic Extraction**
    - Introduction
    - **Conversation Modeling**
    - LeadLDA Topic Model
    - Experiments



[O] Just an hour ago, a series of coordinated *terrorist* *attacks* occurred in *Paris* !!!

[R1] OMG! I can’t believe it’s real. I’ve just been there last month.

[R2] *Gunmen* and *suicide bombers* hit a *concert* hall. More than 100 are *killed* already.

[R3] Oh no! @BonjourMarc R U OK! please reply me for god’s sake

[R5] Don’t worry. I was home.

[R4] My gosh! that sucks:(Poor on u guys…

[R6] poor guys, terrible

[R7] For the safety of *US*, I’m for *Trump* to be *president*, especially after this.

[R8] I repost to support *Donald*. Can’t agree more :-)

[R9]Thanks dude, you’d never regret :-)

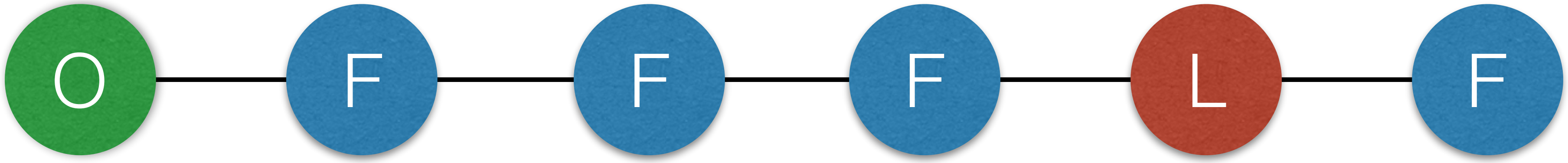
[R10] Are U crazy? *Donald Trump* is just a bigot *sexiest* and *raciest*.

Leaders & Followers

- **Leaders**: raise salient new information
  - Initiate a new topic or
  - Covers important aspects of a previous topic
- **Followers**: echo/respond to parents

Leader Detection

- Simple way: binary classifier on each individual microblog
    - Context information along conversation paths
- Sequence tagging model
    - Conditional Random Field (CRF) (Lafferty et al. 01)



Features for Leader Detection

Feature Category	Feature Description
Text-based	# of terms in m_i
	Part-of-speech of m_i
	Type of sentence of m_i (question or exclamatory)
Microblog-specific	# of emoticons in m_i
	# of hashtags in m_i
	# of urls in m_i
	# of mentions in m_i
Path-specific	Cosine Similarity between m_i and its neighbors
	Cosine Similarity between m_i and root microblog

(m_i denotes the current repost message)

Dataset for Leader Detection

- Data: 1300 conversation paths
  - 1300 original microblogs + 4772 reposts/replies
  - 1000 paths for training and 300 for test
- 3 annotators to label leaders/followers given conversation paths
  - average kappa=0.52 (fair to good agreement)
  - use labels agreed by at least 2 annotators

Leader Detection Evaluation

	Cross-validation			Held-out		
	Precision	Recall	F1	Precision	Recall	F1
Random	0.298	0.495	0.373	0.316	0.496	0.386
LR	0.705	0.663	0.684	0.704	0.662	0.682
SVM	0.709	0.669	0.688	0.689	0.662	0.675
SVMhmm	0.748	0.655	0.698	0.693	0.701	0.697
CRF	**0.755**	**0.720**	**0.737**	**0.711**	**0.707**	**0.709**

Outline

- **Microblog Topic Extraction**
  - Introduction
  - Conversation Modeling
  - **LeadLDA Topic Model**
  - Experiments

Topics and Conversation Trees

- Basic assumptions:
  - One post covers one single topic
  - One conversation tree is a mixture of topics

Topics and Conversation Trees

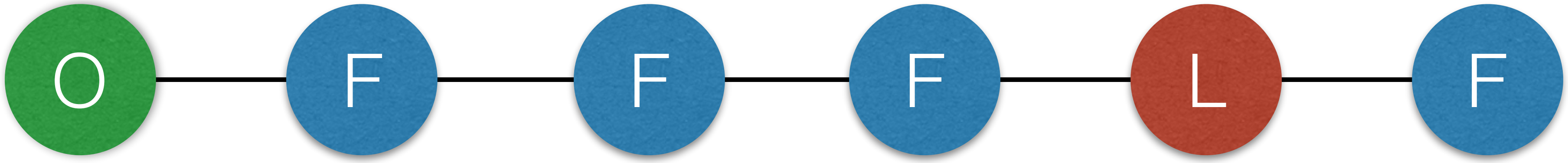
- Topic assignments:
  - Leaders: topic mixture of the conversation tree
  - Followers: the parent-child topic transition
- Word generation:
  - A word being a topical/non-topic word depends on whether it occurs in a leader or a follower post

Prior Information

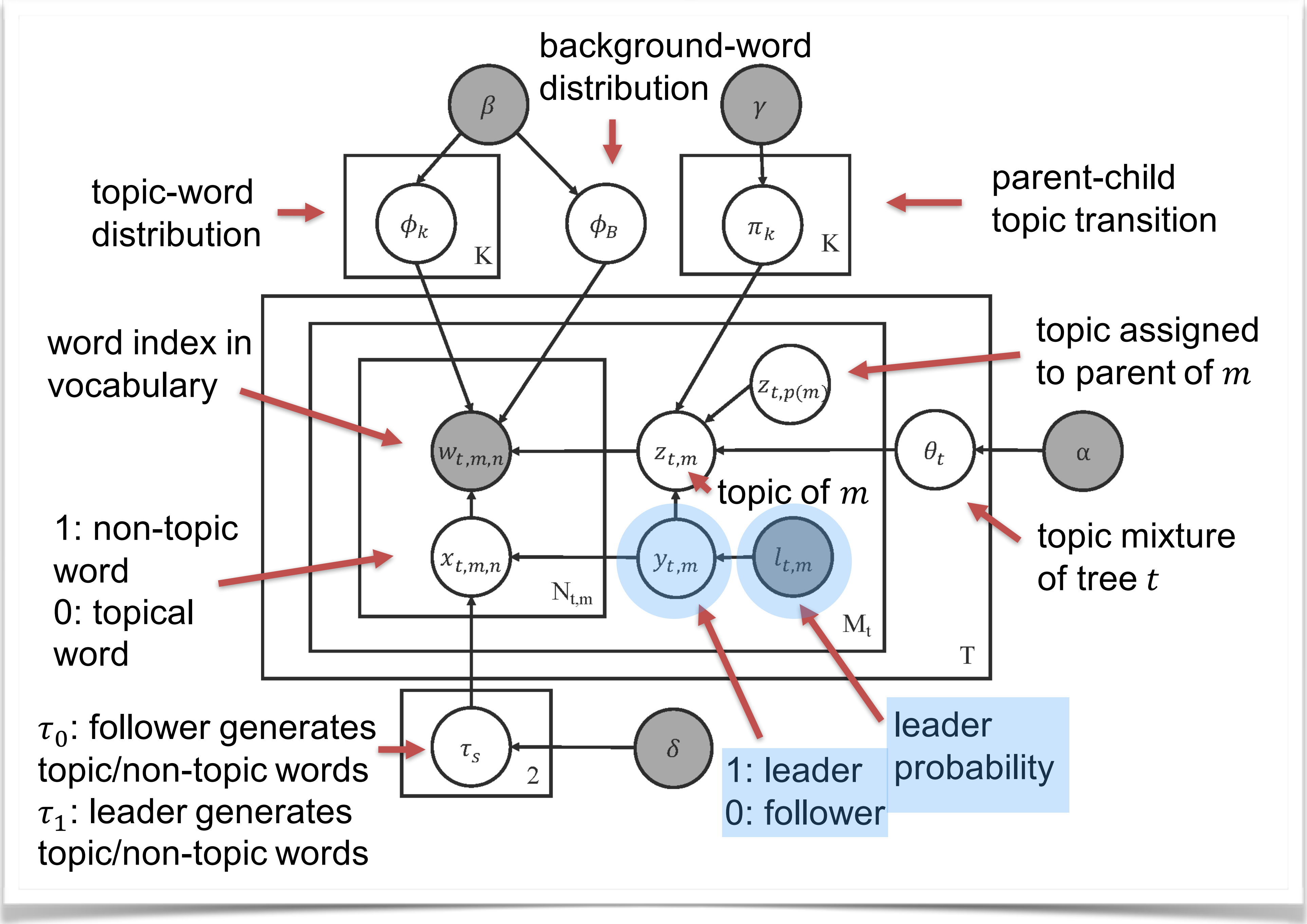
- Leader detection: CRF on conversation paths

- Leader probability: average the marginal probabilities of a same node over all tree paths

- Observed prior variables of the Topic Model



LeadLDA Topic Model



background-word distribution

topic-word distribution

parent-child topic transition

word index in vocabulary

topic assigned to parent of m

topic of m

topic mixture of tree t

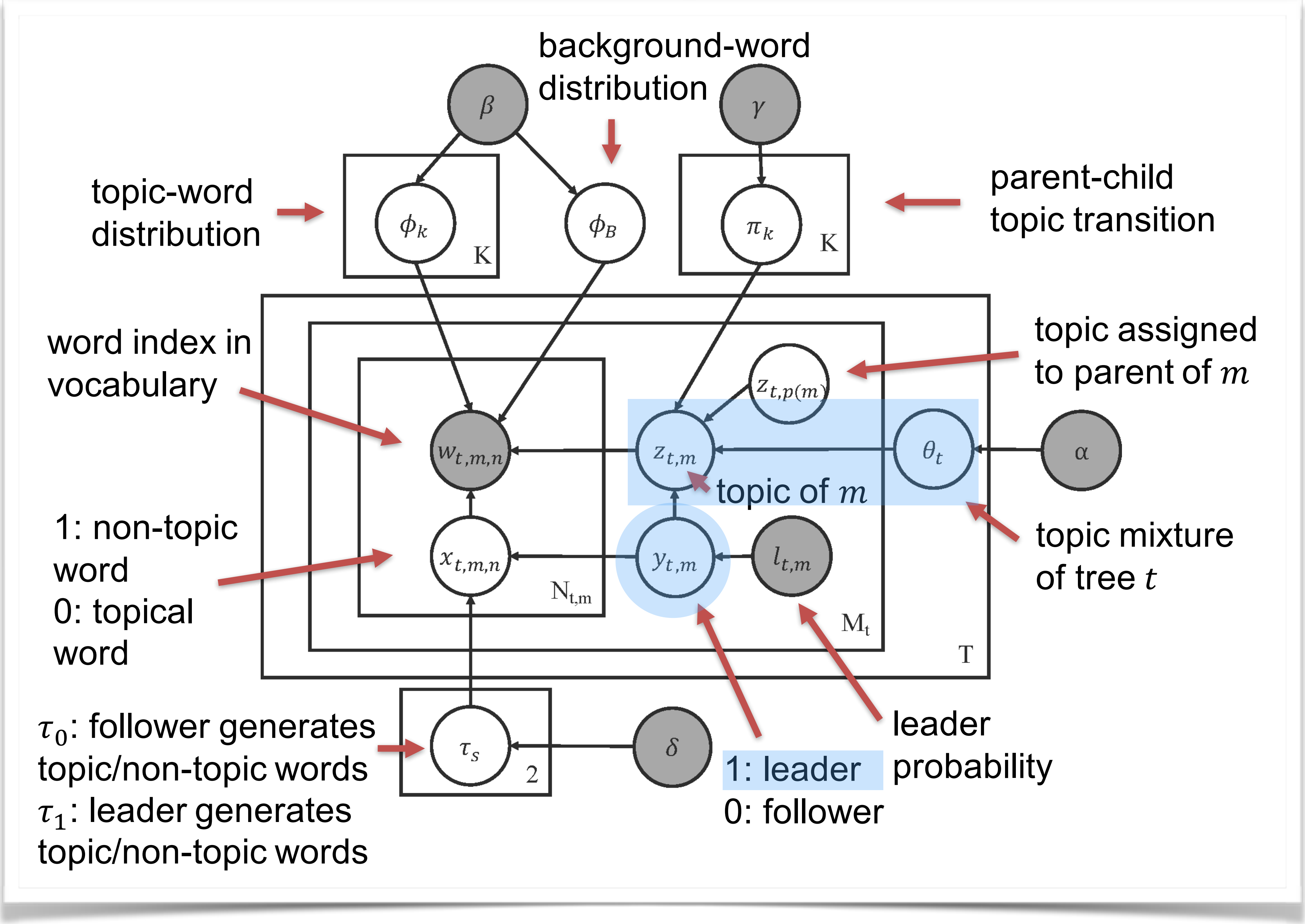
1: non-topic word
0: topical word

τ_0: follower generates topic/non-topic words
τ_1: leader generates topic/non-topic words

1: leader
0: follower

leader probability

LeadLDA Topic Model



background-word distribution

topic-word distribution

parent-child topic transition

word index in vocabulary

topic assigned to parent of m

topic of m

1: non-topic word
0: topical word

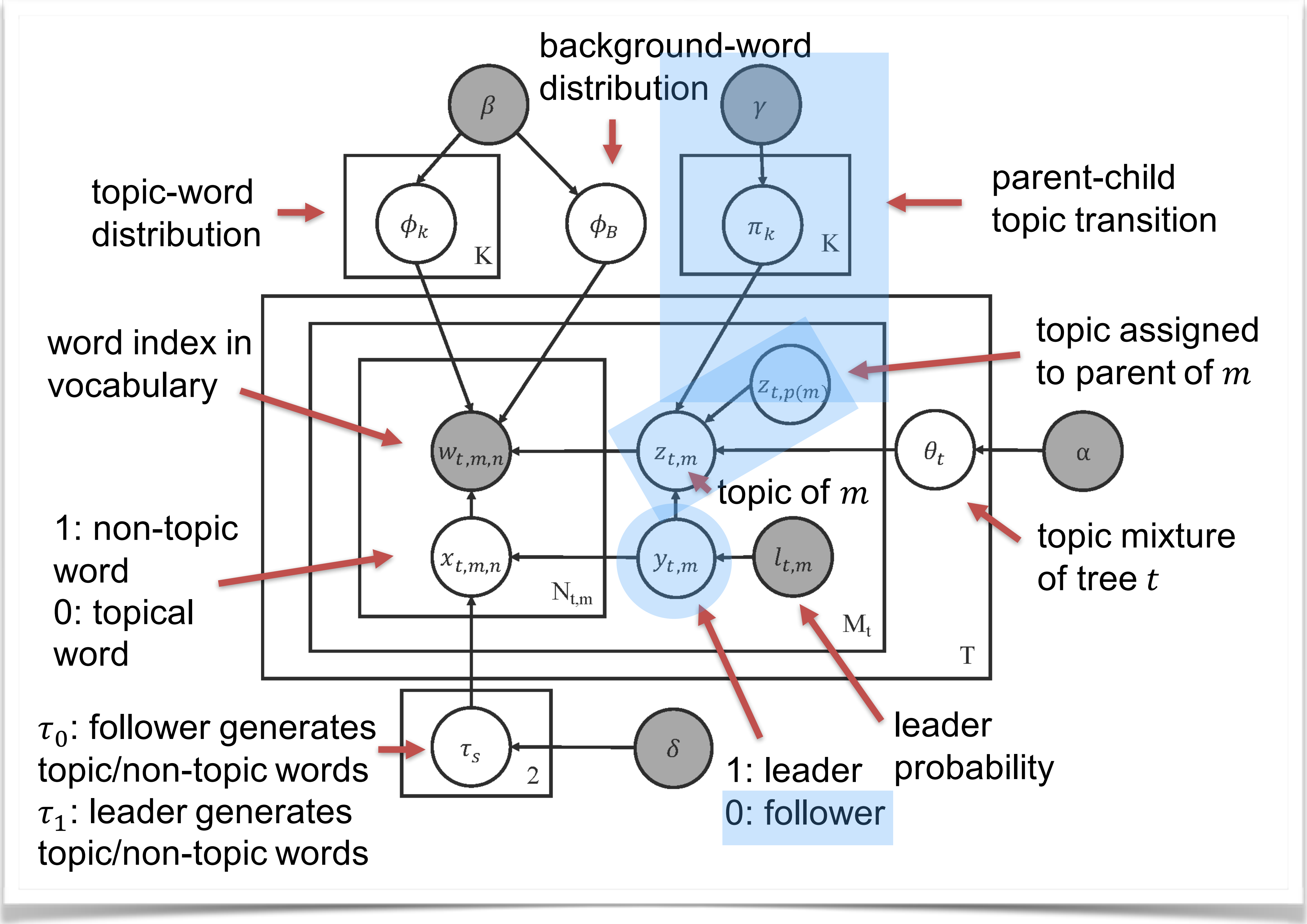
topic mixture of tree t

τ_0: follower generates topic/non-topic words
τ_1: leader generates topic/non-topic words

1: leader
0: follower

leader probability

LeadLDA Topic Model



background-word distribution

topic-word distribution

parent-child topic transition

word index in vocabulary

topic assigned to parent of m

topic of m

1: non-topic word
0: topical word

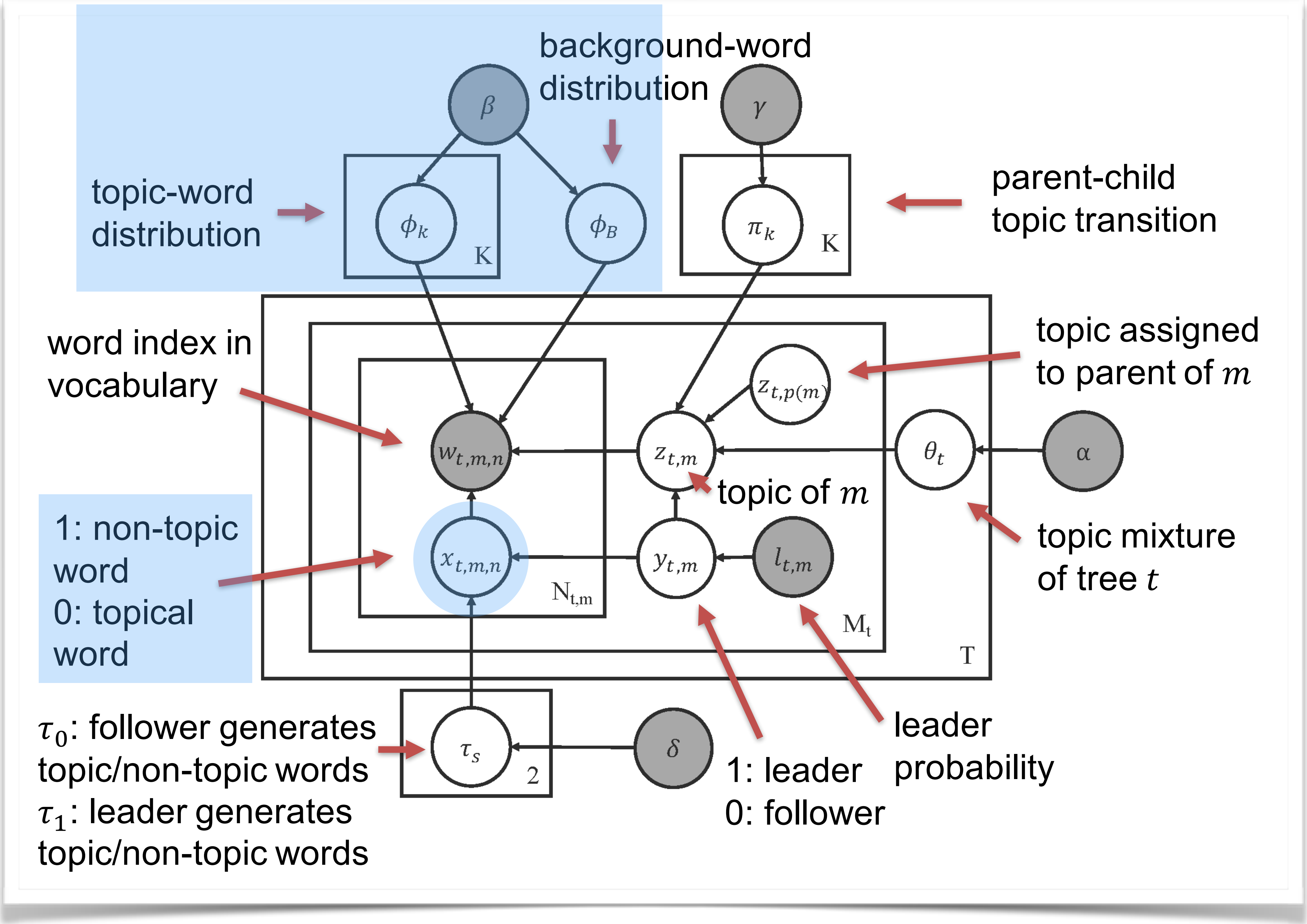
topic mixture of tree t

τ_0: follower generates topic/non-topic words
τ_1: leader generates topic/non-topic words

1: leader
0: follower

leader probability

LeadLDA Topic Model



background-word distribution

topic-word distribution

parent-child topic transition

word index in vocabulary

topic assigned to parent of m

topic of m

1: non-topic word
0: topical word

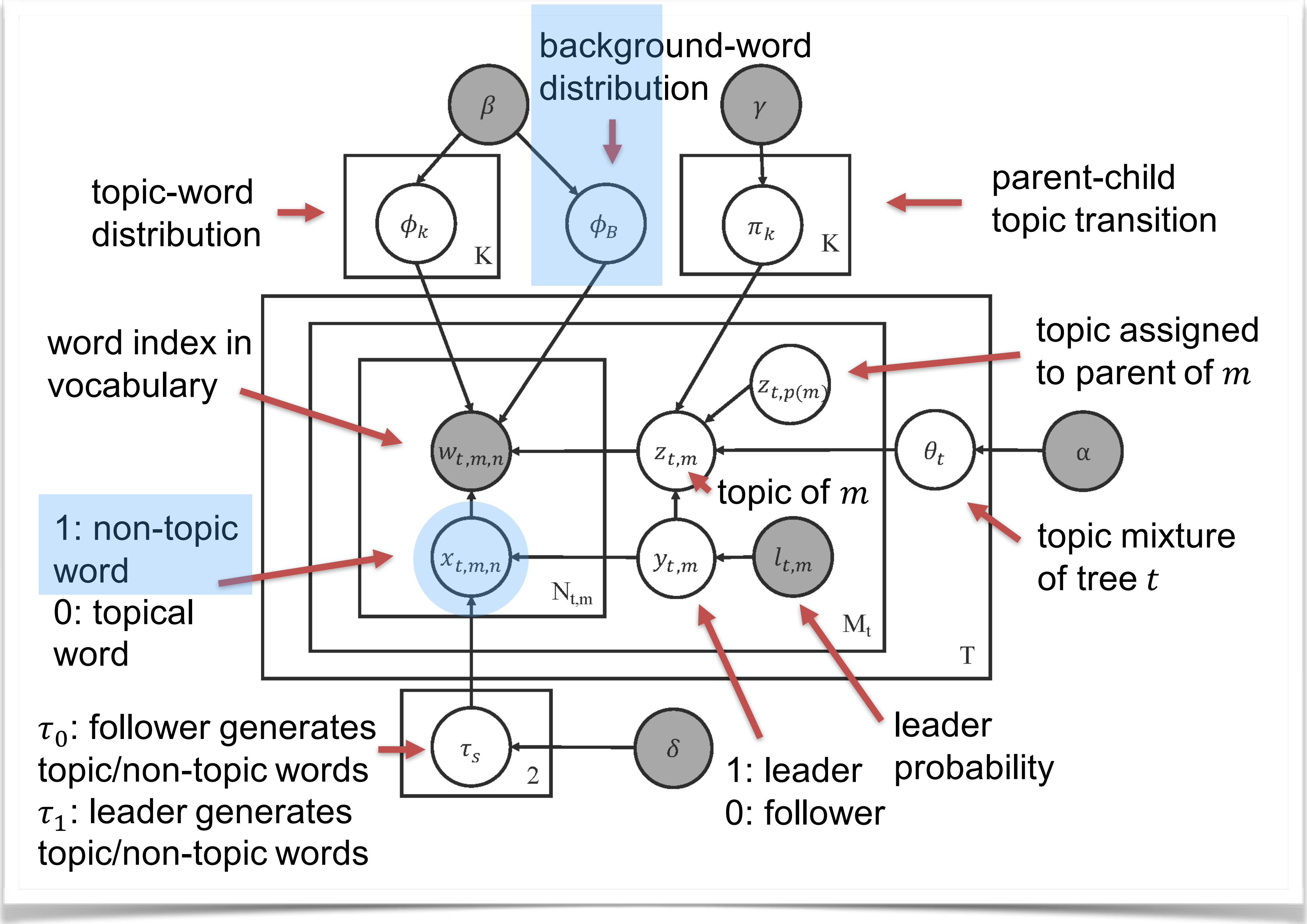
topic mixture of tree t

τ_0: follower generates topic/non-topic words
τ_1: leader generates topic/non-topic words

1: leader
0: follower

leader probability

LeadLDA Topic Model



background-word distribution

topic-word distribution

parent-child topic transition

word index in vocabulary

topic assigned to parent of m

topic of m

1: non-topic word
0: topical word

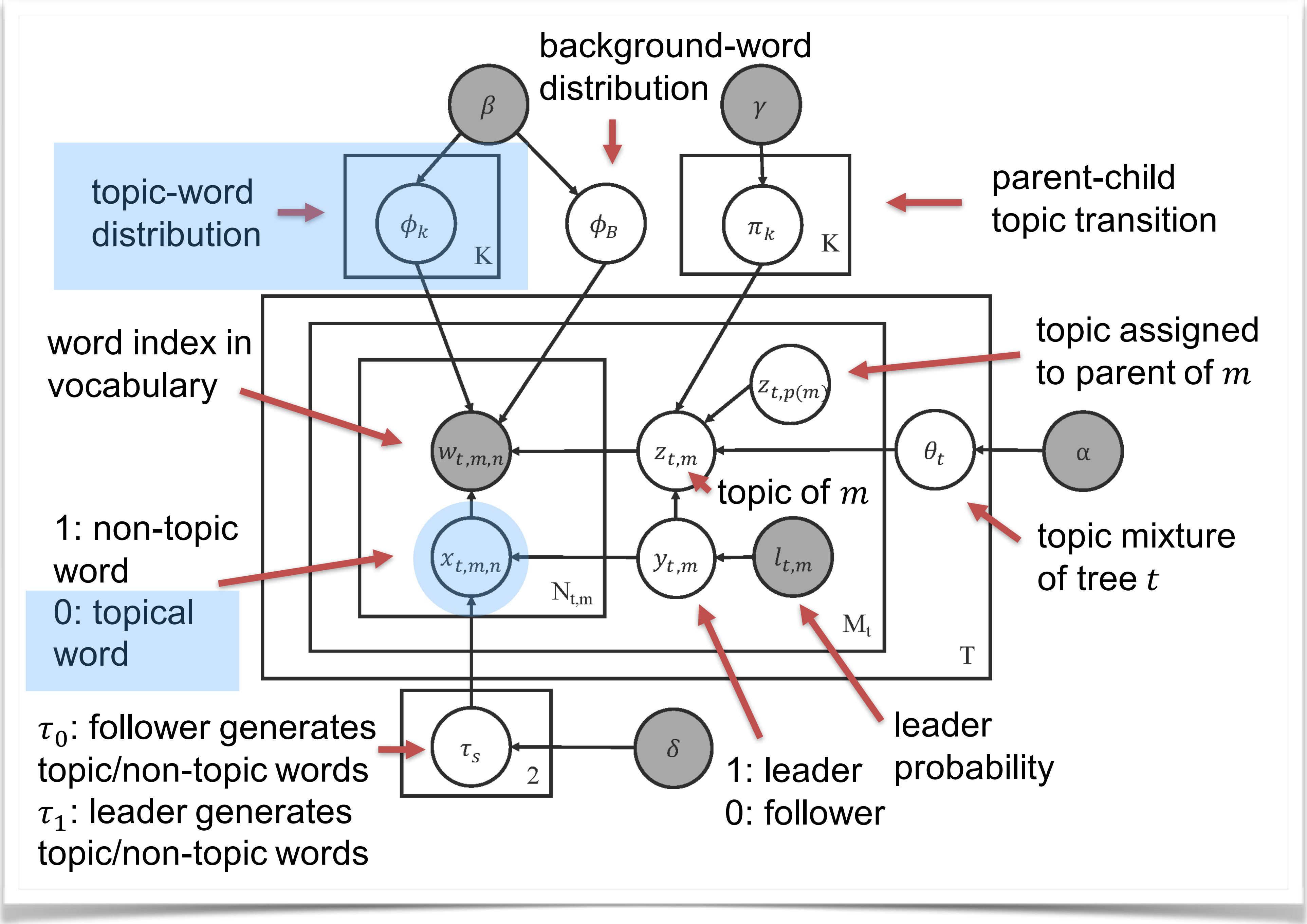
topic mixture of tree t

τ_0: follower generates topic/non-topic words
τ_1: leader generates topic/non-topic words

1: leader
0: follower

leader probability

LeadLDA Topic Model



background-word distribution

topic-word distribution

parent-child topic transition

word index in vocabulary

topic assigned to parent of m

topic of m

1: non-topic word
0: topical word

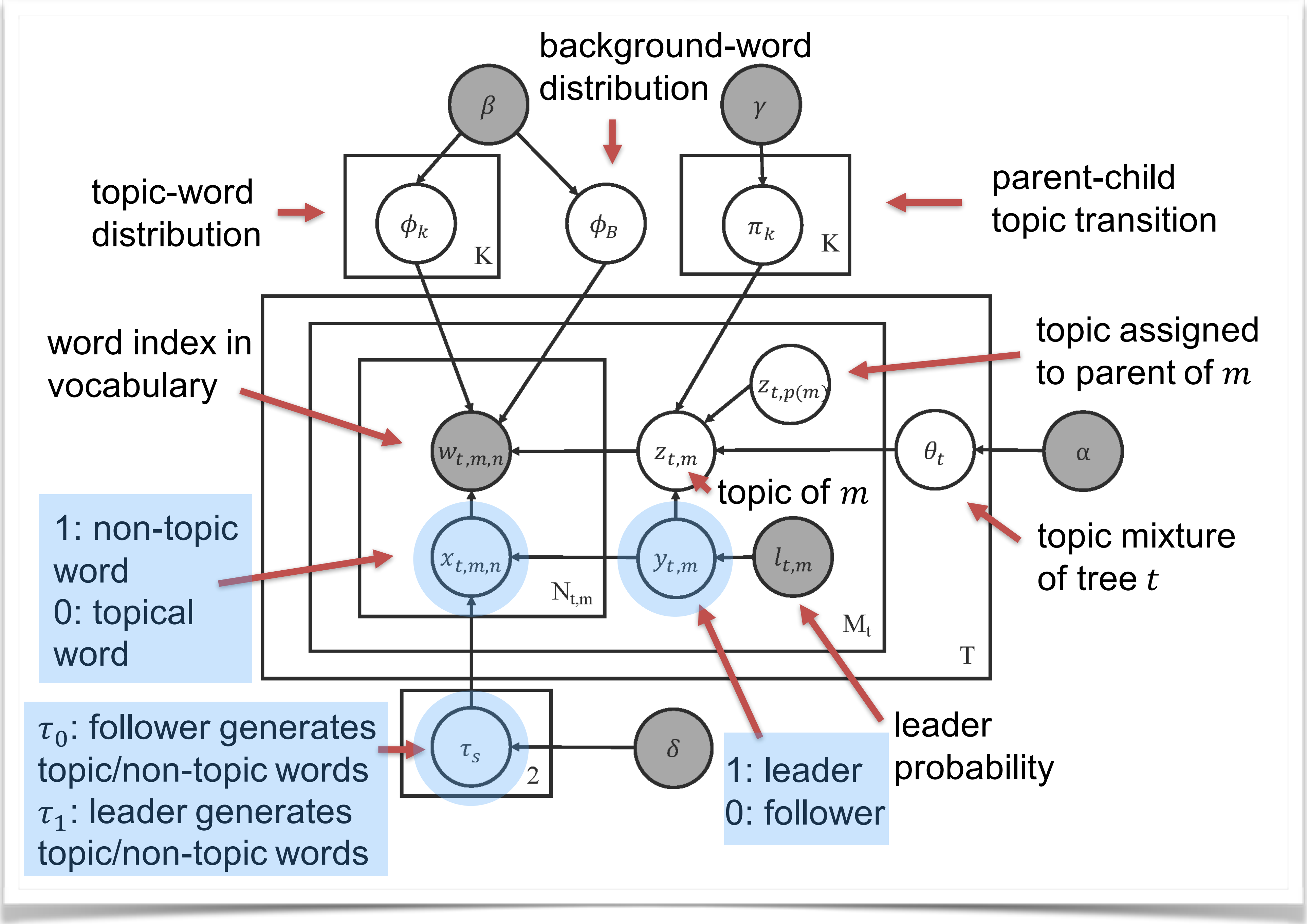
topic mixture of tree t

τ_0: follower generates topic/non-topic words
τ_1: leader generates topic/non-topic words

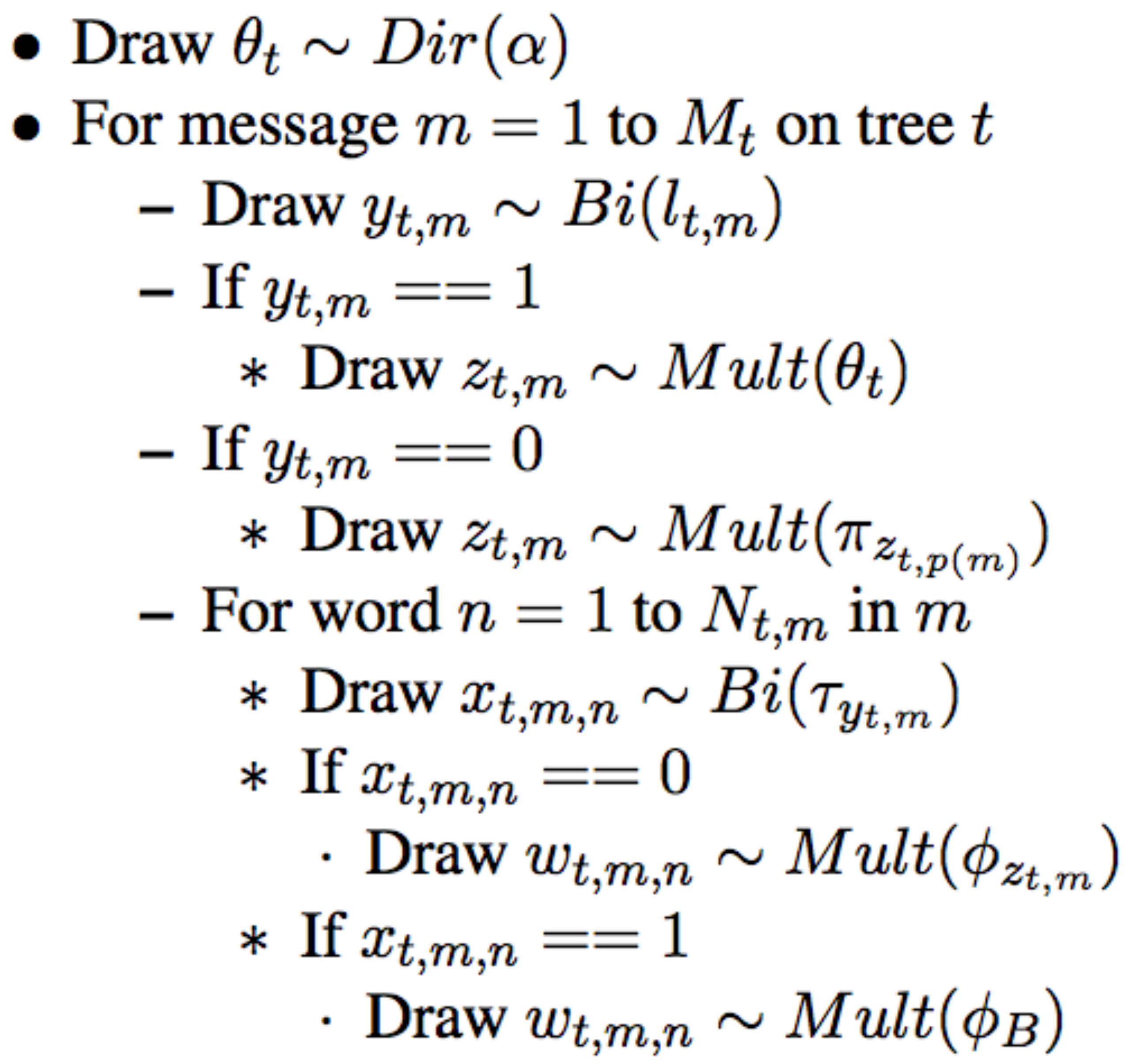
1: leader
0: follower

leader probability

LeadLDA Topic Model

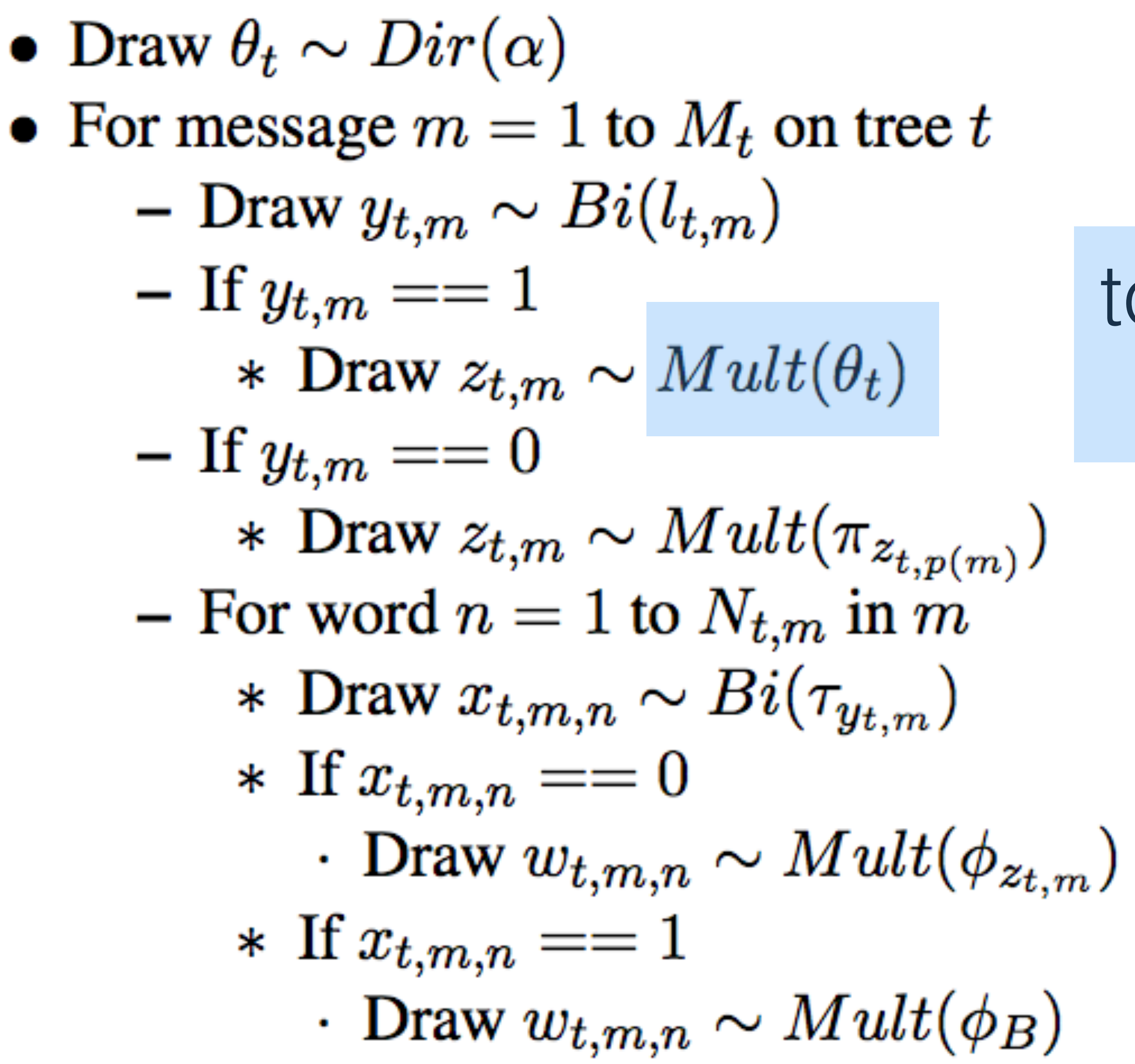


Generation Story



- Draw $\theta_t \sim Dir(\alpha)$
- For message $m = 1$ to M_t on tree t
  - Draw $y_{t,m} \sim Bi(l_{t,m})$
  - If $y_{t,m} == 1$
    - Draw $z_{t,m} \sim Mult(\theta_t)$
  - If $y_{t,m} == 0$
    - Draw $z_{t,m} \sim Mult(\pi_{z_{t,p(m)}})$
  - For word $n = 1$ to $N_{t,m}$ in m
    - Draw $x_{t,m,n} \sim Bi(\tau_{y_{t,m}})$
    - If $x_{t,m,n} == 0$
      - Draw $w_{t,m,n} \sim Mult(\phi_{z_{t,m}})$
    - If $x_{t,m,n} == 1$
      - Draw $w_{t,m,n} \sim Mult(\phi_B)$

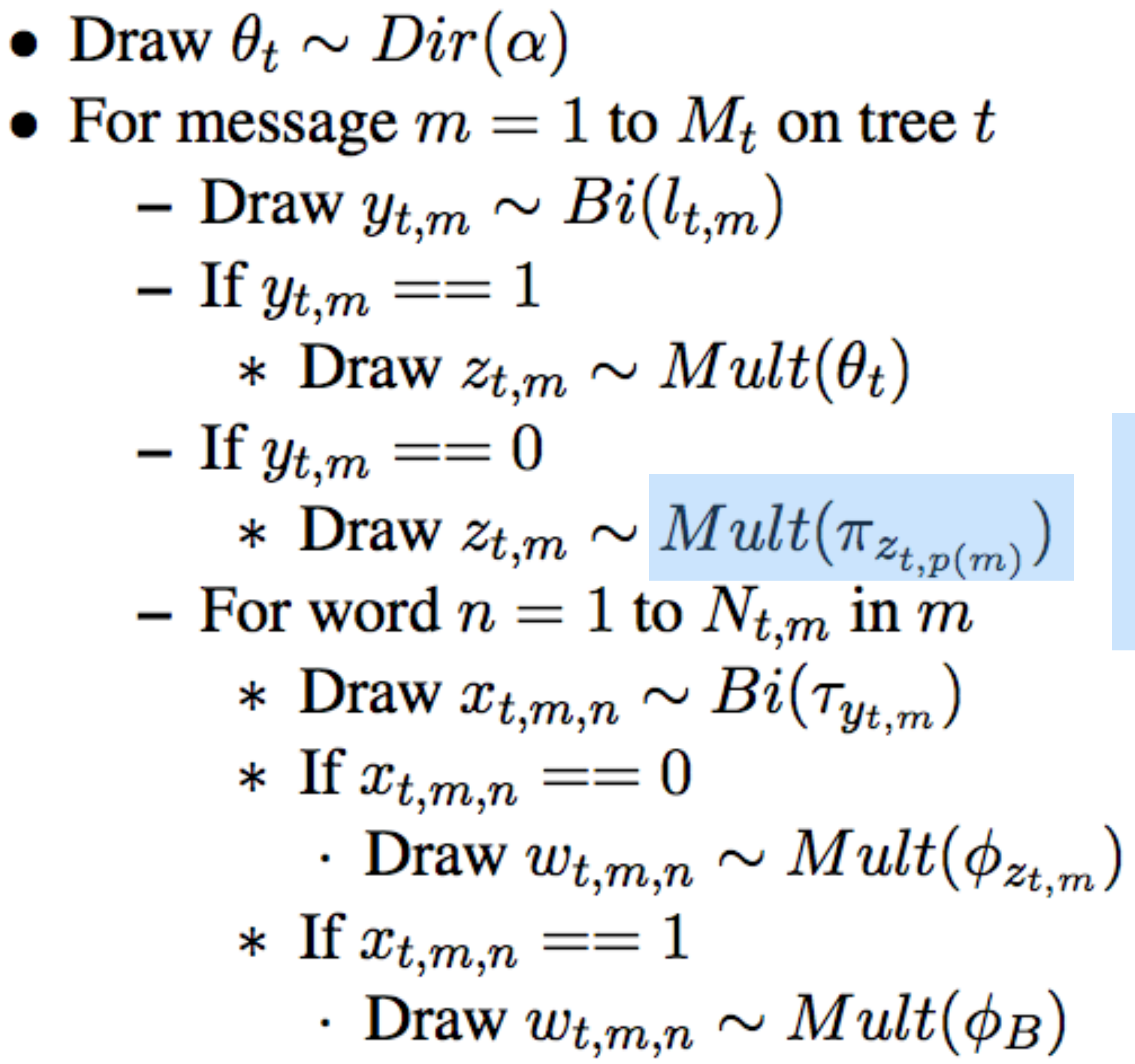
Generation Story



```
• Draw $\theta_t \sim Dir(\alpha)$
• For message $m = 1$ to $M_t$ on tree $t$
    – Draw $y_{t,m} \sim Bi(l_{t,m})$
    – If $y_{t,m} == 1$
        * Draw $z_{t,m} \sim Mult(\theta_t)$
    – If $y_{t,m} == 0$
        * Draw $z_{t,m} \sim Mult(\pi_{z_{t,p(m)}})$
    – For word $n = 1$ to $N_{t,m}$ in $m$
        * Draw $x_{t,m,n} \sim Bi(\tau_{y_{t,m}})$
        * If $x_{t,m,n} == 0$
            · Draw $w_{t,m,n} \sim Mult(\phi_{z_{t,m}})$
        * If $x_{t,m,n} == 1$
            · Draw $w_{t,m,n} \sim Mult(\phi_B)$
```

topic mixture of the conversation tree

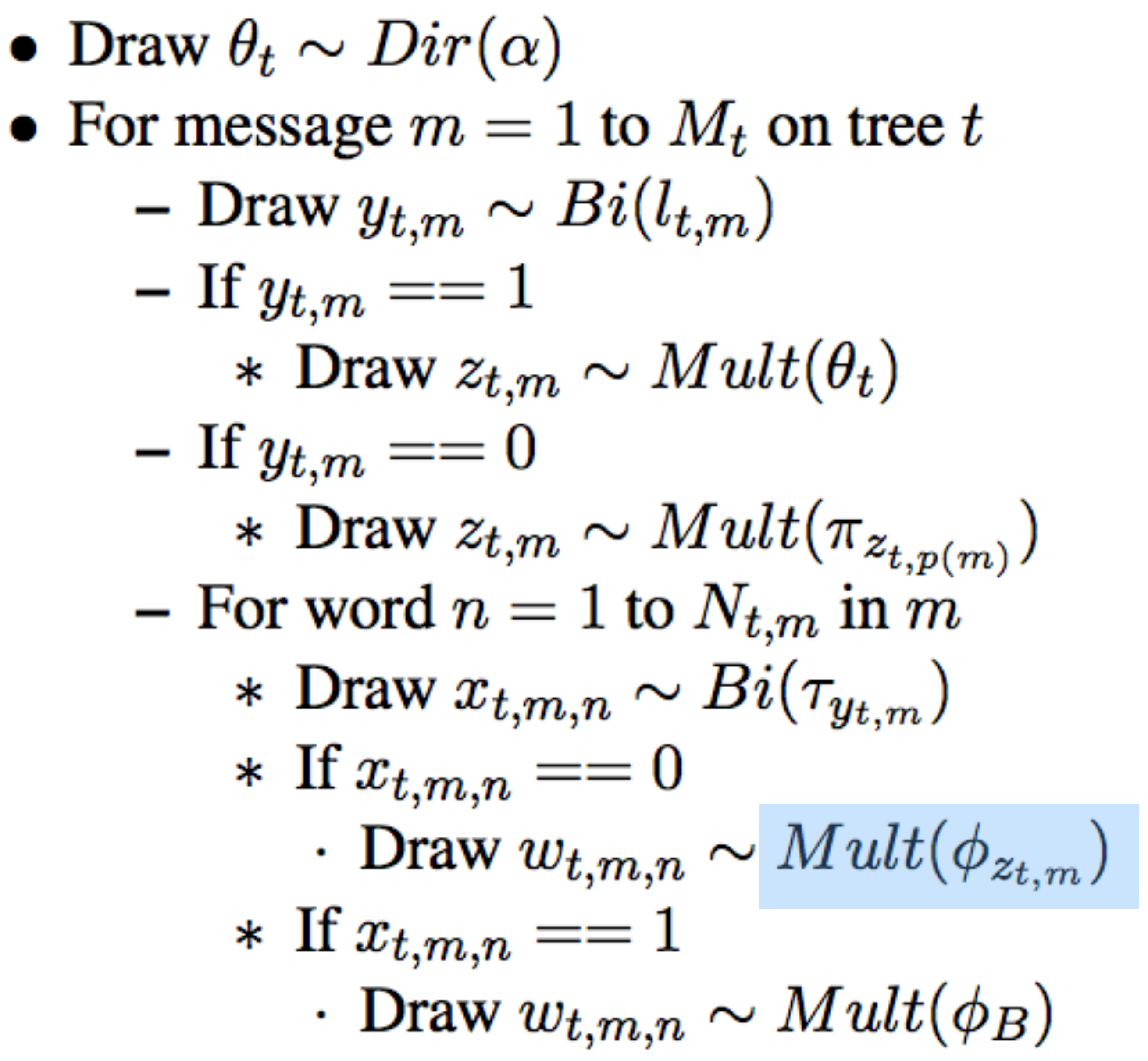
Generation Story



```
• Draw $\theta_t \sim Dir(\alpha)$
• For message $m = 1$ to $M_t$ on tree $t$
    – Draw $y_{t,m} \sim Bi(l_{t,m})$
    – If $y_{t,m} == 1$
        ∗ Draw $z_{t,m} \sim Mult(\theta_t)$
    – If $y_{t,m} == 0$
        ∗ Draw $z_{t,m} \sim Mult(\pi_{z_{t,p(m)}})$
    – For word $n = 1$ to $N_{t,m}$ in $m$
        ∗ Draw $x_{t,m,n} \sim Bi(\tau_{y_{t,m}})$
        ∗ If $x_{t,m,n} == 0$
            · Draw $w_{t,m,n} \sim Mult(\phi_{z_{t,m}})$
        ∗ If $x_{t,m,n} == 1$
            · Draw $w_{t,m,n} \sim Mult(\phi_B)$
```

parent-child topic transitions

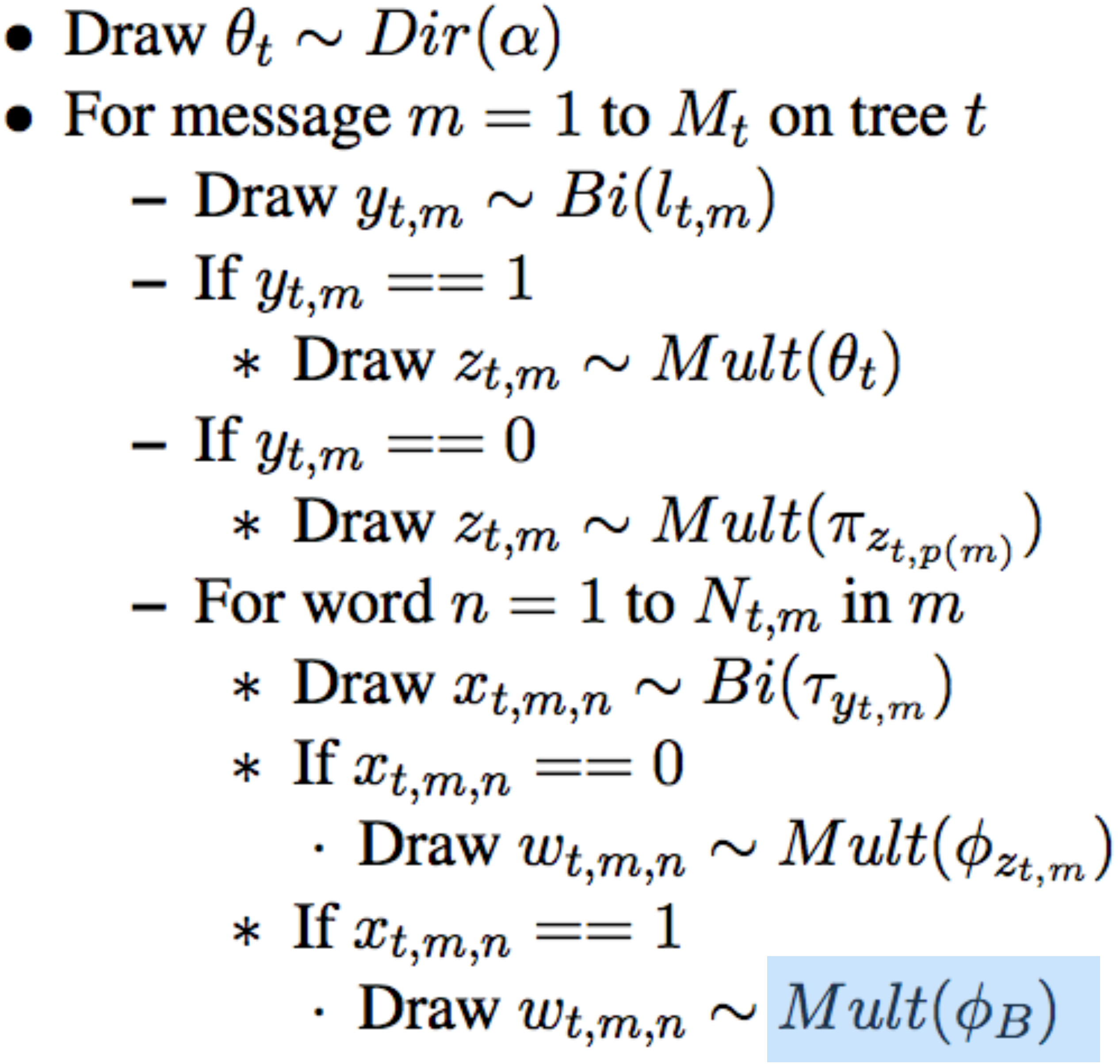
Generation Story



- Draw $\theta_t \sim Dir(\alpha)$
- For message $m = 1$ to M_t on tree t
  - Draw $y_{t,m} \sim Bi(l_{t,m})$
  - If $y_{t,m} == 1$
    - Draw $z_{t,m} \sim Mult(\theta_t)$
  - If $y_{t,m} == 0$
    - Draw $z_{t,m} \sim Mult(\pi_{z_{t,p(m)}})$
  - For word $n = 1$ to $N_{t,m}$ in m
    - Draw $x_{t,m,n} \sim Bi(\tau_{y_{t,m}})$
    - If $x_{t,m,n} == 0$
      - Draw $w_{t,m,n} \sim Mult(\phi_{z_{t,m}})$
    - If $x_{t,m,n} == 1$
      - Draw $w_{t,m,n} \sim Mult(\phi_B)$

topic-word distribution

Generation Story



```
• Draw $\theta_t \sim Dir(\alpha)$
• For message $m = 1$ to $M_t$ on tree $t$
    – Draw $y_{t,m} \sim Bi(l_{t,m})$
    – If $y_{t,m} == 1$
        ∗ Draw $z_{t,m} \sim Mult(\theta_t)$
    – If $y_{t,m} == 0$
        ∗ Draw $z_{t,m} \sim Mult(\pi_{z_{t,p(m)}})$
    – For word $n = 1$ to $N_{t,m}$ in $m$
        ∗ Draw $x_{t,m,n} \sim Bi(\tau_{y_{t,m}})$
        ∗ If $x_{t,m,n} == 0$
            · Draw $w_{t,m,n} \sim Mult(\phi_{z_{t,m}})$
        ∗ If $x_{t,m,n} == 1$
            · Draw $w_{t,m,n} \sim Mult(\phi_B)$
```

background-word distribution

Inference

- Collapsed Gibbs Sampling
- The hidden multinomial variables

leader switcher

  - Message-level: y z topic assignment
  - Word-level: x background switcher

- are sampled in turn conditioned on a complete assignment of all other hidden variables.

Outline

- **Microblog Topic Extraction**
  - Introduction
  - Conversation Modeling
  - LeadLDA Topic Model
  - **Experiments**

Evaluation Datasets

Month	# of trees	# of msgs	Vocab size
May	10,812	38,926	6,011
June	29,547	98,001	9,539
July	26,103	102,670	10,121

Baselines

- TreeLDA: all posts are assumed to be leaders
- StructLDA: all posts are assumed to be followers
- BTM: directly model topics of biterms in posts
- SATM: jointly aggregate posts and infer topics

Objective Analysis

- Coherence scores:
  - $$\mathcal{C} = \frac{1}{K} \cdot \sum_{k=1}^{K} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{D(w_i^k, w_j^k) + 1}{D(w_j^k)}$$
- Words representing a coherent topic are likely to co-occur within the same “document”
- Documents: microblog posts tagged by the same hashtag

Objective Analysis

Model	May		June		July	
	K50	K100	K50	K100	K50	K100
TREE	-138.8	-138.6	-102.0	-115.0	-115.8	-119.7
STR	-134.0	-136.9	-104.3	-112.7	-111.0	-117.3
BTM	-125.2	-131.1	-109.4	-115.7	-115.3	-120.2
SATM	-134.6	-131.9	-105.5	-114.3	-113.5	-118.9
LEAD	**-120.9**	**-127.2**	**-101.6**	**-106.0**	**-97.2**	**-104.9**

Subjective Analysis

Model	May		June		July	
	K50	K100	K50	K100	K50	K100
TREE	3.12	3.41	3.42	3.44	3.03	3.48
STR	3.05	3.45	3.38	3.48	3.08	3.53
BTM	3.04	3.26	3.40	3.37	3.15	3.57
SATM	3.08	3.43	3.30	3.55	3.09	3.54
LEAD	**3.40**	**3.57**	**3.52**	**3.63**	**3.55**	**3.72**

Case Study

TreeLDA	StructLDA	BTM	SATM	LeadLDA
香港 **微博** 马航 家属 证实 入境处 客机 消息 **曹格 投给 二胎 选项 教父 滋养** 飞机 外国 心情 坠毁 男子 同胞	乌克兰 航空 **亲爱** 国民 绕开 飞行 航班 领空 所有 避开 宣布 空域 东部 俄罗斯 终于 **忘记** 公司 绝望 **看看** 珍贵	香港 入境处 家属 证实 男子 护照 外国 消息 坠毁 马航 报道 联系 电台 客机 飞机 同胞 确认 事件 **霍家** 直接	马航 祈祷 安息 生命 逝者 世界 艾滋病 恐怖 **广州** 飞机 无辜 默哀 远离 事件 击落 **公交车** 中国人 国际 愿逝者 真的	乌克兰 马航 客机 **击落** 飞机 坠毁 **导弹** 俄罗斯 消息 乘客 中国 马来西亚 香港 遇难 事件 武装 航班 恐怖 目前 证实
Hong Kong, **microblog**, Malaysia Airlines, family, confirm, immigration, airliner, news, **Grey Chow**, **vote**, **second baby**, **choice**, **god father**, **nourish**, airplane, foreign, feeling, crash, man,	Ukraine, airline, **dear**, national, bypass, fly, flight, airspace, all, avoid, announce, airspace, eastern Russia, finally, **forget**, company, disappointed, **look**, valuable	Hong Kong, immigration, family, confirm, man, passport, foreign, news, crash, Malaysia Airlines, report, contact, broadcast station, airliner, airplane, fellowman, confirm, event, **Fok's family**, directly	Malaysia Airlines, prey, rest in peace, life, dead, world, AIDS, terror, **Guangzhou**, airplane, innocent, silent tribute, keep away from, event, shoot down, **bus**, Chinese, international, wish the dead, really	Ukraine, Malaysia Airlines, airliner, **shoot down**, airplane, crash, **missile**, Russia, news, passenger, China, Malaysia, Hong Kong, killed, event, militant, flight, terror, current, confirm

Outline

- Background
- Microblog Topic Extraction
- **Conversation Tree Summarization**
- Conclusion

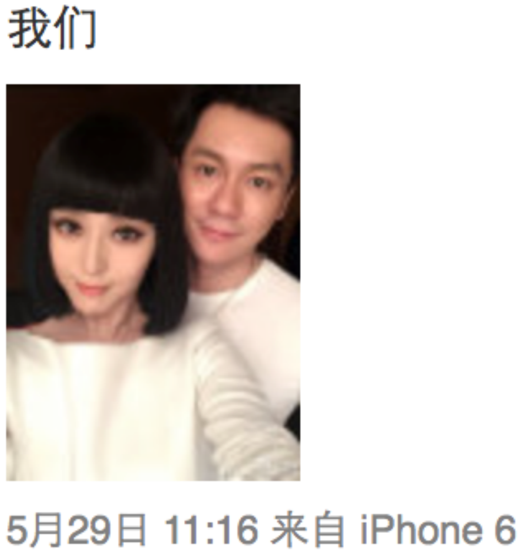
Outline

- **Conversation Tree Summarization**
  - **Introduction**
  - LeadSum Summarization Model
  - Experiments

Introduction

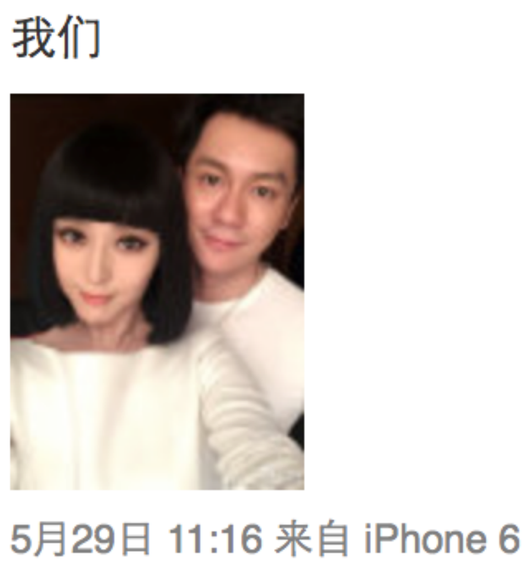
- An individual microblog message is short and lack of context information
    - Cannot capture key information of a message
    - E.g., a message advertising iPhone 7
- Reposts/replies provide valuable context information to a microblog

李晨： 我们
LI, Chen: We



我们

5月29日 11:16 来自 iPhone 6

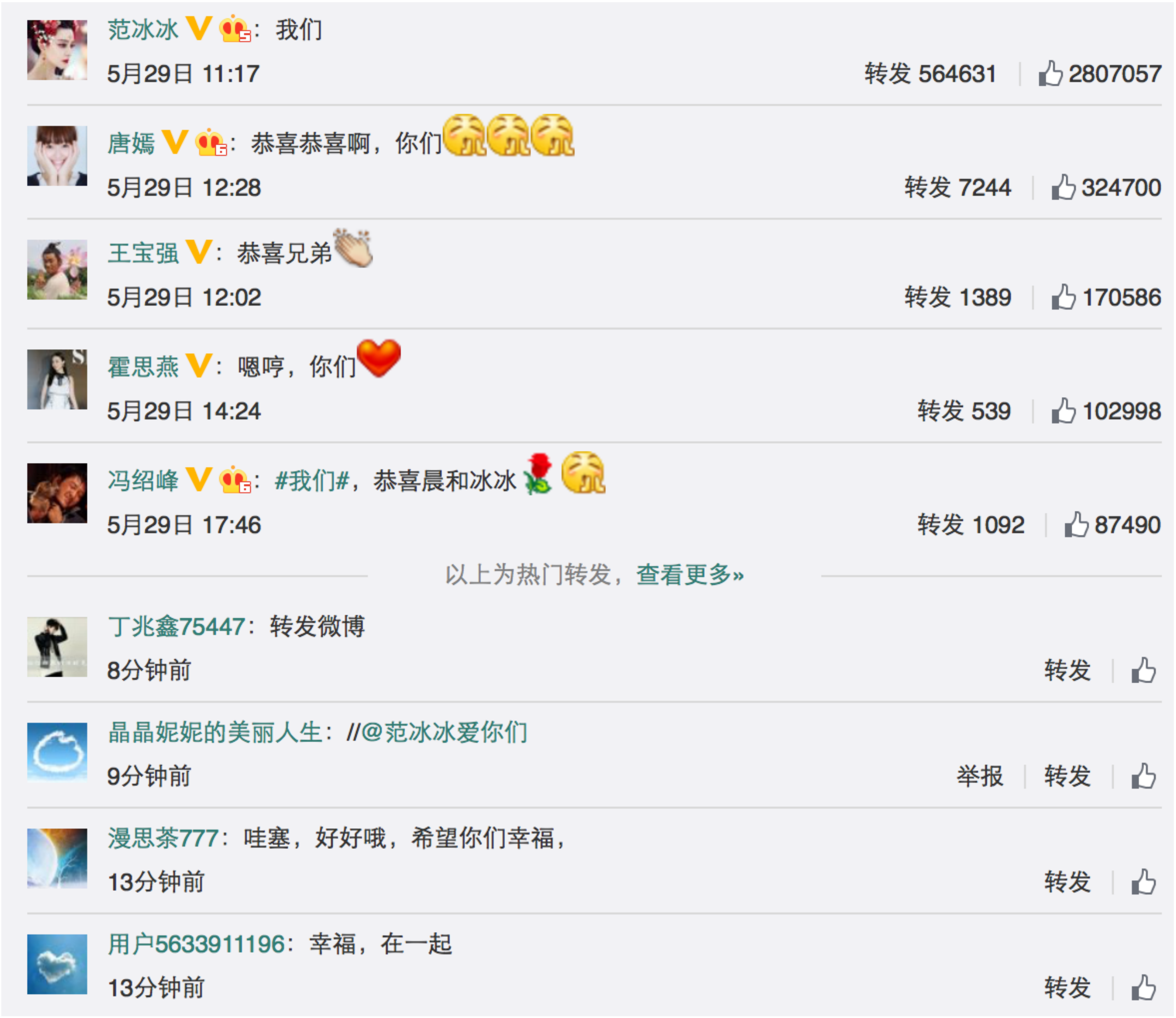


我们

5月29日 11:16 来自 iPhone 6

Original Microblog

Reposts/Replies



范冰冰: 我们
5月29日 11:17 转发 564631 | 2807057

唐嫣: 恭喜恭喜啊，你们
5月29日 12:28 转发 7244 | 324700

王宝强: 恭喜兄弟
5月29日 12:02 转发 1389 | 170586

霍思燕: 嗯哼，你们❤
5月29日 14:24 转发 539 | 102998

冯绍峰: #我们#，恭喜晨和冰冰
5月29日 17:46 转发 1092 | 87490

以上为热门转发，查看更多»

丁兆鑫75447: 转发微博
8分钟前 转发 |

晶晶妮妮的美丽人生: //@范冰冰爱你们
9分钟前 举报 | 转发 |

漫思茶777: 哇塞，好好哦，希望你们幸福，
13分钟前 转发 |

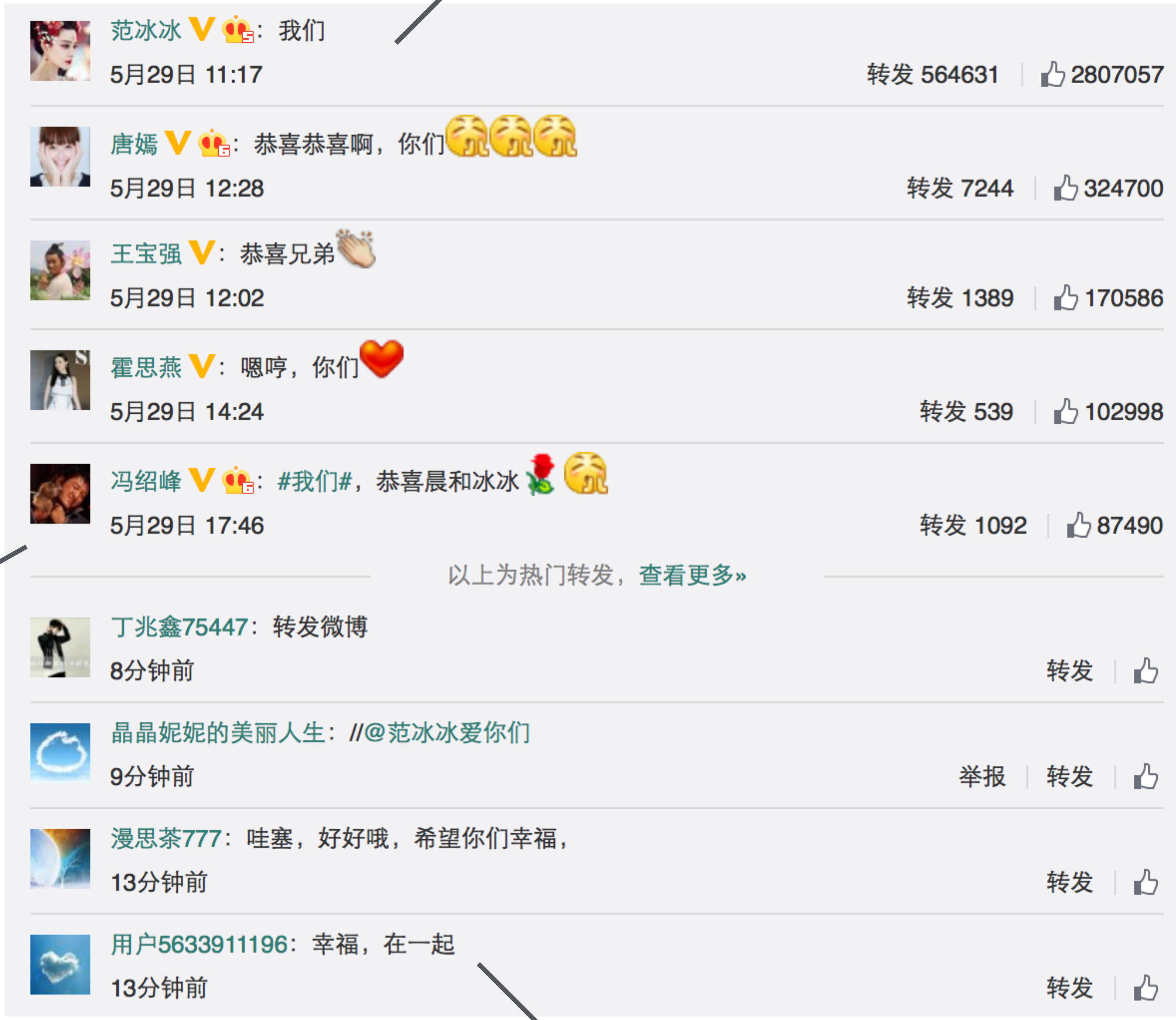
用户5633911196: 幸福，在一起
13分钟前 转发 |

李晨：我们
LI Chen: We

我们

5月29日 11:16 来自 iPhone 6

范冰冰：我們
FAN, Bingbing: We



范冰冰 V：我们
5月29日 11:17 转发 564631 | 2807057

唐嫣 V：恭喜恭喜啊，你们
5月29日 12:28 转发 7244 | 324700

王宝强 V：恭喜兄弟
5月29日 12:02 转发 1389 | 170586

霍思燕 V：嗯哼，你们❤
5月29日 14:24 转发 539 | 102998

冯绍峰 V：#我们#，恭喜晨和冰冰🌹
5月29日 17:46 转发 1092 | 87490

以上为热门转发，查看更多»

丁兆鑫75447：转发微博
8分钟前 转发 |

晶晶妮妮的美丽人生：//@范冰冰爱你们
9分钟前 举报 | 转发 |

漫思茶777：哇塞，好好哦，希望你们幸福，
13分钟前 转发 |

用户5633911196：幸福，在一起
13分钟前 转发 |

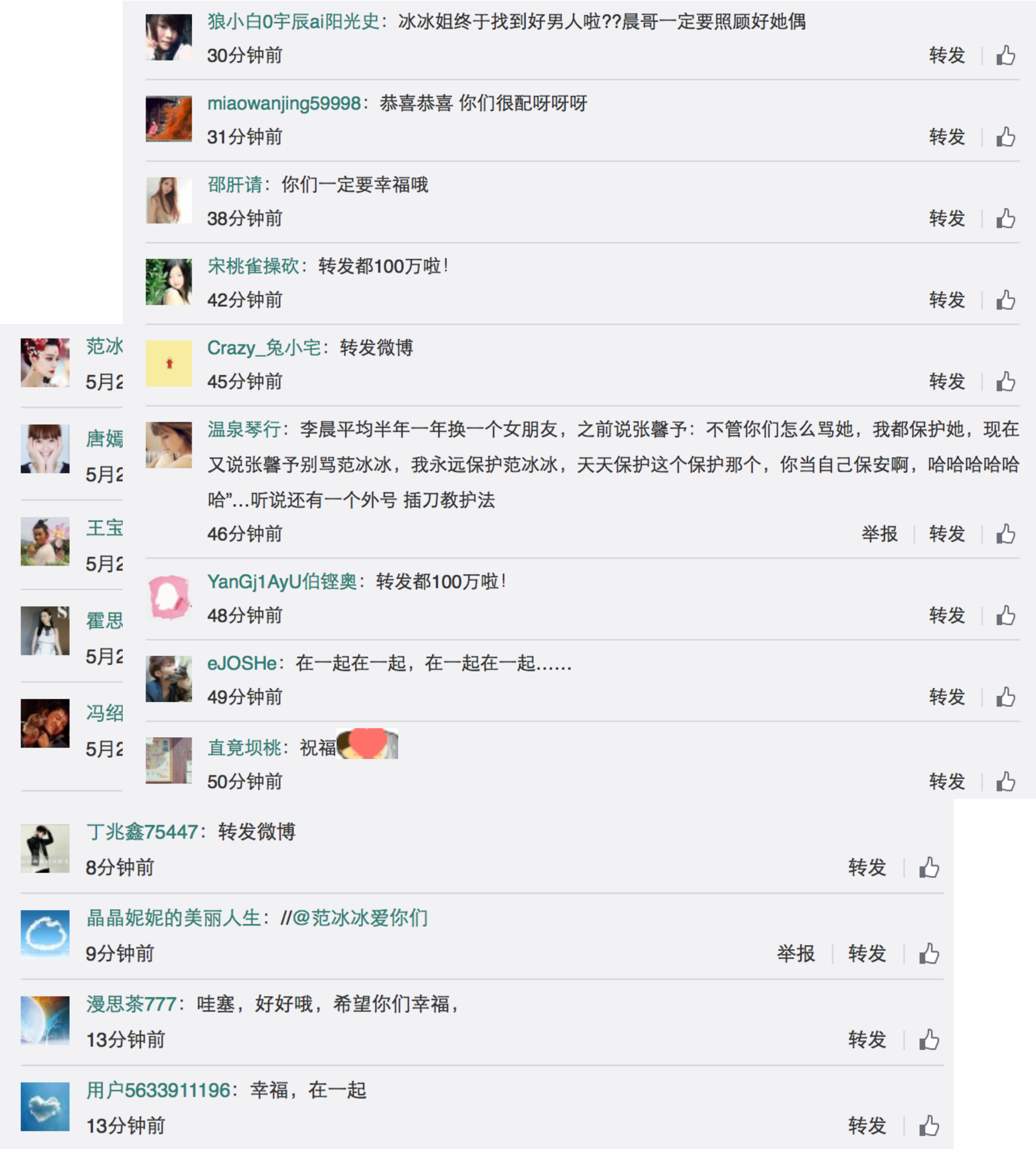
冯绍峰：恭喜晨和冰冰
FENG, Shaofeng: Congrats to Chen and Bingbing

用户5***6：幸福，在一起
User5***6: Sweet love

李晨： 我们
LI Chen: We

我们

5月29日 11:16 来自 iPhone 6



狼小白0宇辰ai阳光史：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
30分钟前 转发

miaowanjing59998：恭喜恭喜 你们很配呀呀呀
31分钟前 转发

邵肝请：你们一定要幸福哦
38分钟前 转发

宋桃雀操砍：转发都100万啦！
42分钟前 转发

Crazy_兔小宅：转发微博
45分钟前 转发

温泉琴行：李晨平均半年一年换一个女朋友，之前说张馨予：不管你们怎么骂她，我都保护她，现在又说张馨予别骂范冰冰，我永远保护范冰冰，天天保护这个保护那个，你当自己保安啊，哈哈哈哈哈哈"...听说还有一个外号 插刀教护法
46分钟前 举报 转发

YanGj1AyU佰铿奥：转发都100万啦！
48分钟前 转发

eJOSHe：在一起在一起，在一起在一起......
49分钟前 转发

直竟坝桃：祝福
50分钟前 转发

范冰
5月2

唐嫣
5月2

王宝
5月2

霍思
5月2

冯绍
5月2

丁兆鑫75447：转发微博
8分钟前 转发

晶晶妮妮的美丽人生：//@范冰冰爱你们
9分钟前 举报 转发

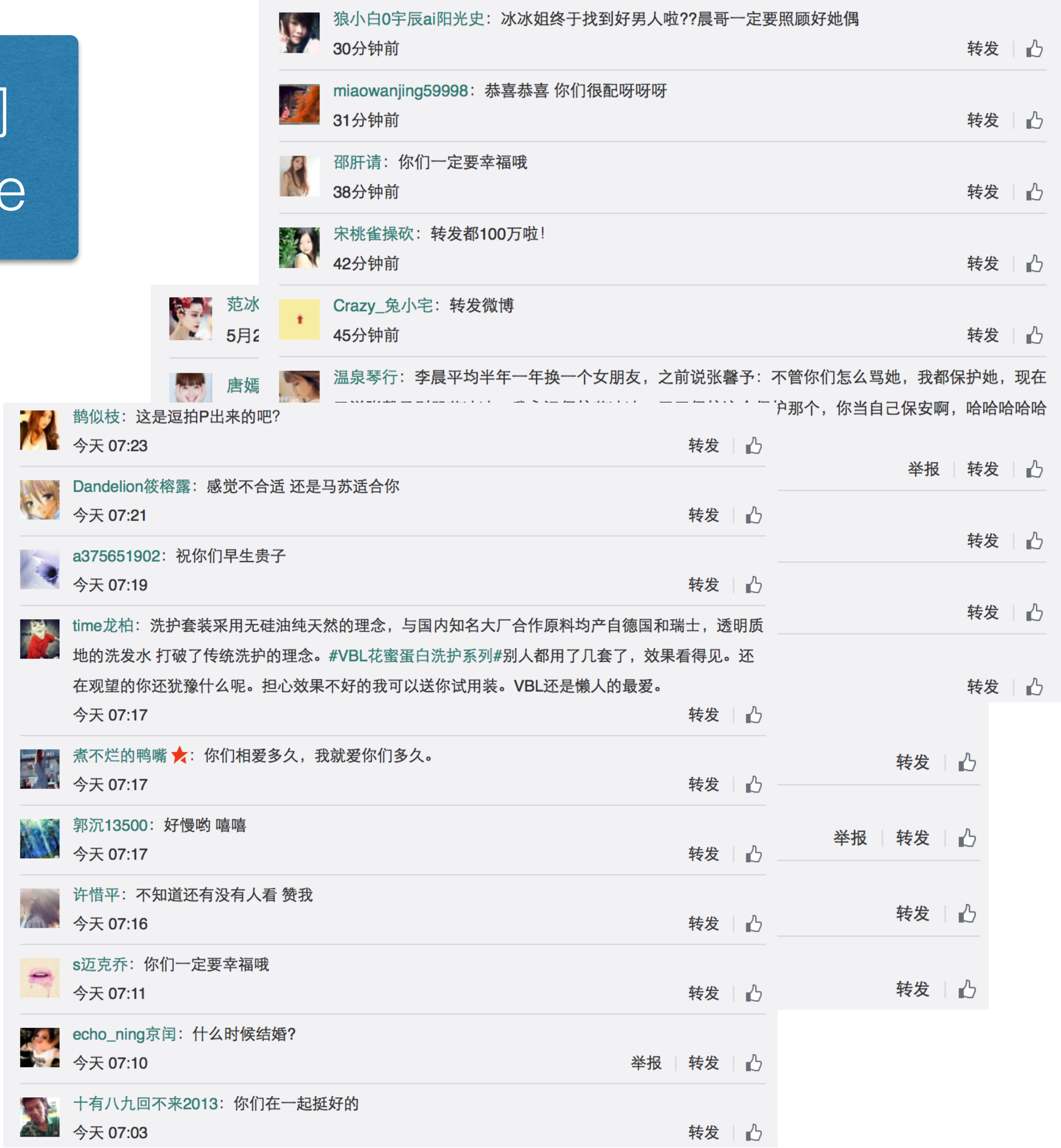
漫思茶777：哇塞，好好哦，希望你们幸福，
13分钟前 转发

用户5633911196：幸福，在一起
13分钟前 转发

李晨：我们
LI Chen: We

我们

5月29日 11:16 来自 iPhone 6



狼小白0宇辰ai阳光史：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
30分钟前 转发

miaowanjing59998：恭喜恭喜 你们很配呀呀呀
31分钟前 转发

邵肝请：你们一定要幸福哦
38分钟前 转发

宋桃雀操砍：转发都100万啦！
42分钟前 转发

Crazy_兔小宅：转发微博
45分钟前 转发

温泉琴行：李晨平均半年一年换一个女朋友，之前说张馨予：不管你们怎么骂她，我都保护她，现在
护那个，你当自己保安啊，哈哈哈哈哈
举报 转发

范冰
5月2

唐嫣

鹊似枝：这是逗拍P出来的吧?
今天 07:23 转发

Dandelion筱榕露：感觉不合适 还是马苏适合你
今天 07:21 转发

a375651902：祝你们早生贵子
今天 07:19 转发

time龙柏：洗护套装采用无硅油纯天然的理念，与国内知名大厂合作原料均产自德国和瑞士，透明质地的洗发水 打破了传统洗护的理念。#VBL花蜜蛋白洗护系列#别人都用了几套了，效果看得见。还在观望的你还犹豫什么呢。担心效果不好的我可以送你试用装。VBL还是懒人的最爱。
今天 07:17 转发

煮不烂的鸭嘴★：你们相爱多久，我就爱你们多久。
今天 07:17 转发

郭沉13500：好慢哟 嘻嘻
今天 07:17 转发

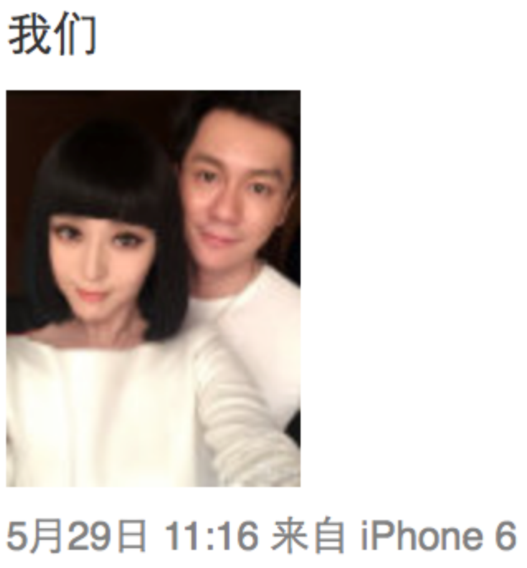
许惜平：不知道还有没有人看 赞我
今天 07:16 转发

s迈克乔：你们一定要幸福哦
今天 07:11 转发

echo_ning京闰：什么时候结婚?
今天 07:10 举报 转发

十有八九回不来2013：你们在一起挺好的
今天 07:03 转发

李晨：我们
LI Chen: We



我们

5月29日 11:16 来自 iPhone 6

狼小白0宇辰ai阳光史：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
30分钟前 转发

miaowanjing59998：恭喜恭喜 你们很配呀呀呀
31分钟前 转发

邵肝请：你们一定要幸福哦
38分钟前 转发

宋桃雀操砍：转发都100万啦！
42分钟前 转发

Crazy_兔小宅：转发微博
45分钟前 转发

温泉琴行：李晨平均半年一年换一个女朋友，之前说张馨予：不管你们怎么骂她，我都保护她，现在
户那个，你当自己保安啊，哈哈哈哈哈
举报 转发

转发

转发

转发

范冰
5月2

唐嫣

鹊似枝：这是逗拍P出来的吧?
今天 07:23 转发

Dandelion筱榕露：感觉不合适 还是马苏适合你
今天 07:21 转发

a375651902：祝你们早生贵子
今天 07:19 转发

time龙柏：洗护套装采用无硅油纯天然的理念，与国内知名大厂合作原料均产自德国和瑞士，透明质地的洗发水 打破了传统洗护的理念。#VBL花蜜蛋白洗护系列#别人都用了几套了，效果看得见。还在观望的你还犹豫什么呢。担心效果不好的我可以送你试用装。VBL还是懒人的最爱。
转发

：你们相爱多久，我就爱你们多久。
转发

慢哟 嘻嘻
转发

还有没有人看 赞我
转发

定要幸福哦
转发

什么时候结婚?
举报 转发

2013：你们在一起挺好的
转发

转发

举报 转发

转发

转发

嘘你看你看你看不见：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
今天 04:34 转发

happy孙楠楠：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
今天 04:29 转发

芳在我心一片：好慢哟 嘻嘻
今天 04:29 转发

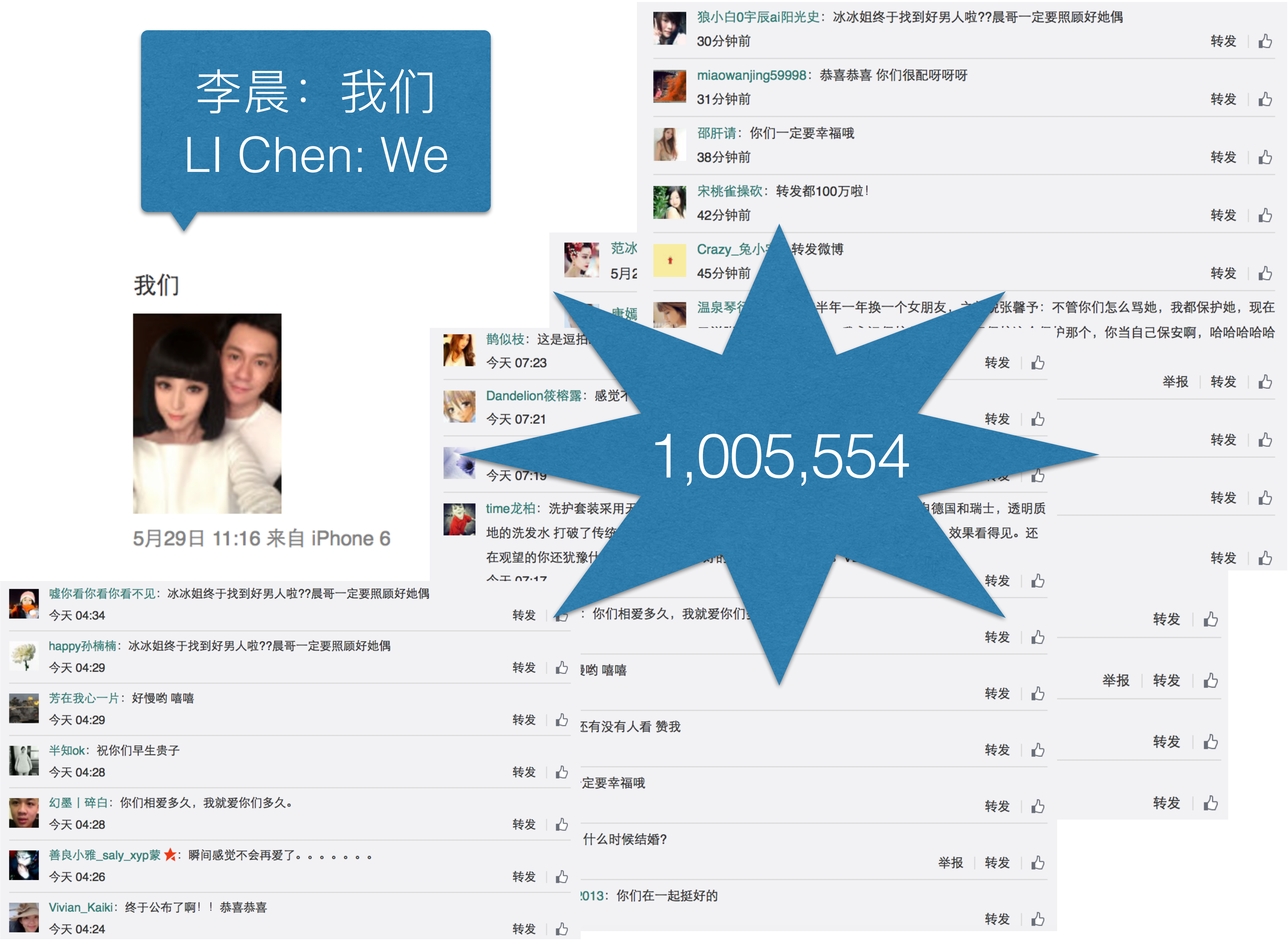
半知ok：祝你们早生贵子
今天 04:28 转发

幻墨丨碎白：你们相爱多久，我就爱你们多久。
今天 04:28 转发

善良小雅_saly_xyp蒙★：瞬间感觉不会再爱了。。。。。。。
今天 04:26 转发

Vivian_Kaiki：终于公布了啊！！恭喜恭喜
今天 04:24 转发

李晨： 我们
LI Chen: We



我们

5月29日 11:16 来自 iPhone 6

1,005,554

狼小白0宇辰ai阳光史：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
30分钟前 转发

miaowanjing59998：恭喜恭喜 你们很配呀呀呀
31分钟前 转发

邵肝请：你们一定要幸福哦
38分钟前 转发

宋桃雀操砍：转发都100万啦！
42分钟前 转发

Crazy_兔小 转发微博
45分钟前 转发

温泉琴 半年一年换一个女朋友， 张馨予：不管你们怎么骂她，我都保护她，现在 护那个，你当自己保安啊，哈哈哈哈哈
举报 转发

转发

转发

转发

范冰
5月2

唐嫣

鹊似枝：这是逗拍
今天 07:23 转发

Dandelion筱榕露：感觉不
今天 07:21 转发

今天 07:19

time龙柏：洗护套装采用天 德国和瑞士，透明质
地的洗发水 打破了传统 效果看得见。还
在观望的你还犹豫什
今天 07:17 转发

：你们相爱多久，我就爱你们多
转发

哟 嘻嘻
转发

还有没有人看 赞我
转发

定要幸福哦
转发

什么时候结婚?
举报 转发

2013：你们在一起挺好的
转发

转发

举报 转发

转发

转发

嘘你看你看你看不见：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
今天 04:34 转发

happy孙楠楠：冰冰姐终于找到好男人啦??晨哥一定要照顾好她偶
今天 04:29 转发

芳在我心一片：好慢哟 嘻嘻
今天 04:29 转发

半知ok：祝你们早生贵子
今天 04:28 转发

幻墨丨碎白：你们相爱多久，我就爱你们多久。
今天 04:28 转发

善良小雅_saly_xyp蒙★：瞬间感觉不会再爱了。。。。。。。
今天 04:26 转发

Vivian_Kaiki：终于公布了啊！！恭喜恭喜
今天 04:24 转发

Microblog Context Summarization

- Problem definition (Chang et al. 2013):
  - Input: an original microblog + all its reposts/replies
  - Output: a succinct summary with a small subset of reposts/replies
- An intuitive solution:
  - Directly apply conventional extractive summarizers
  - Microblog posts are short and informal rendering the lack of context information in each individual messages

Prior Works

- Twitter context tree summarization (Chang et al. 2013)
  - Treat reposts as tweet streams
  - Utilize GBDT model to rank and summarize reposts
    - Author interaction features
    - Need tremendous external historical user interaction data
    - Reposts/replies of influential users might not be salient summary candidates necessarily

Our Idea

- Resort to conversation structures
  - Enrich contextual information
  - Provide clues to identify salient messages for summarization

Leaders & Followers

- **Leaders**: raise salient new information
  - lead further discussions in descendants
  - represent the key content in followers
- **Followers**: echo/respond to parents
  - less important than followers

Outline

- **Conversation Tree Summarization**
- Introduction
- **LeadSum Summarization Model**
- Experiments

Basic-LeadSum Model

- DivRank: random walk based ranking model that balance high information coverage and low redundancy in top ranking vertices. (Mei et al. 10)

- To reduce noise in summary: select messages only from leader messages.

$$p_t(u \to v) = (1-\mu) \cdot p_0(v) + \mu \cdot \frac{p_0(u \to v) N_{t-1}(v))}{\sum_{w \in V_L} p_0(u \to w) N_{t-1}(w)}$$

Basic-LeadSum Model

- DivRank

- To reduce noise in summary: select messages only from leader messages.

sim(u,v)

The times visitor visits v

$$p_t(u \rightarrow v) = (1-\mu) \cdot p_0(v) + \mu \cdot \frac{p_0(u \rightarrow v)N_{t-1}(v))}{\sum_{w \in V_L} p_0(u \rightarrow w)N_{t-1}(w)}$$

$$\frac{1}{|V_L|}$$

Basic-LeadSum Model

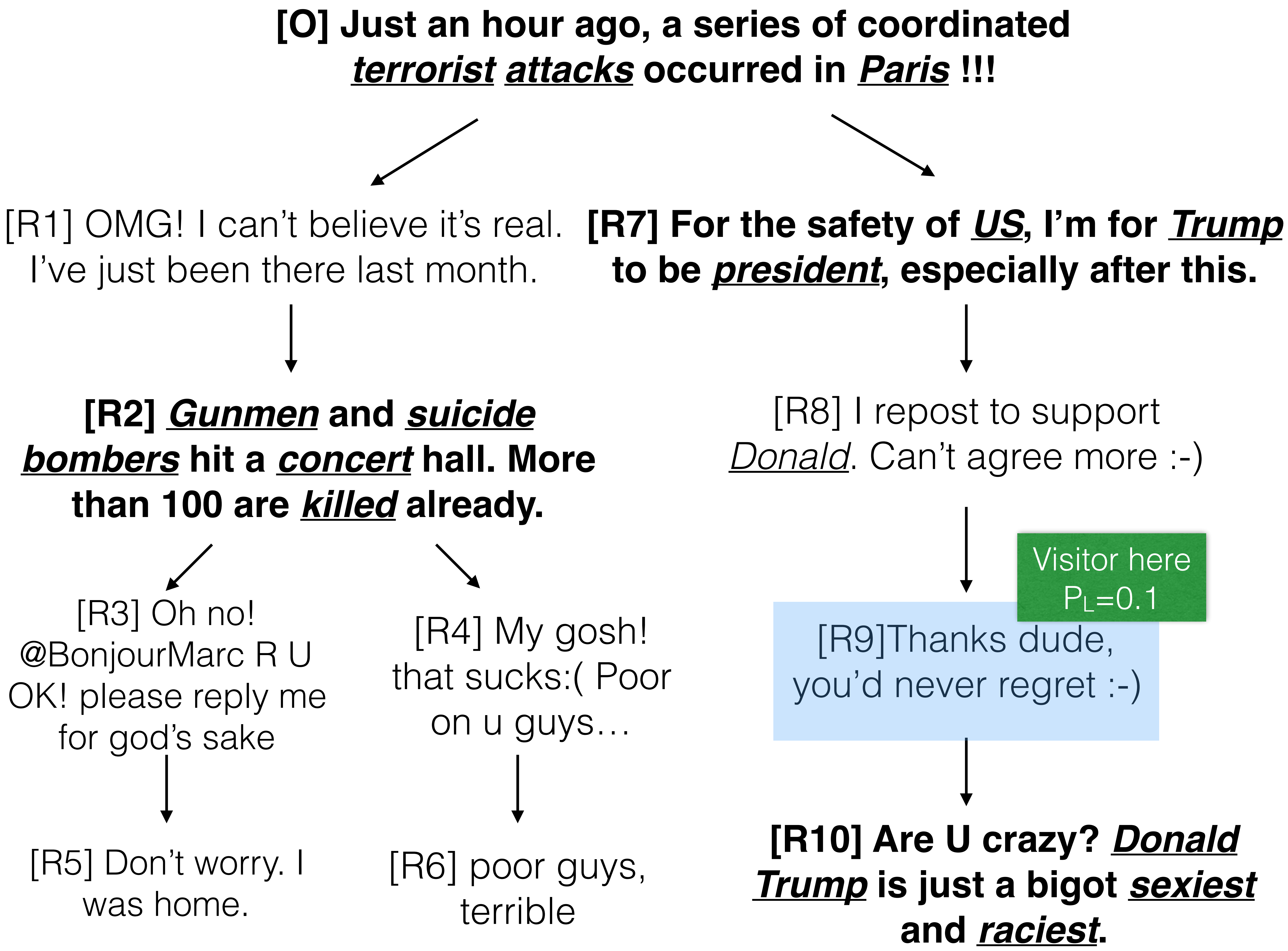
- Error propagation from leader detection model
  - Leaders misclassified as followers (False Negative): leave out strong summary candidates
  - Followers misidentified as leaders (False Positive): may extract real followers in to summary

Soft-LeadSum Model

- Soft-LeadSum: even-length random walk model
    - Let all reposts to participate in summary ranking that reduces FN
    - WALK-1: Transition probability of DivRank
    - WALK-2: a sampling process based on leader probability to avoid selecting real followers

Soft-LeadSum Model

- Soft-LeadSum: even-length random walk model
  - Let all reposts to participate in summary ranking that reduces FN
  - WALK-1: Transition probability of DivRank
  - WALK-2: a sampling process based on leader probability to avoid selecting real followers



[O] Just an hour ago, a series of coordinated *terrorist attacks* occurred in *Paris* !!!

[R1] OMG! I can't believe it's real. I've just been there last month.

[R2] *Gunmen* and *suicide bombers* hit a *concert* hall. More than 100 are *killed* already.

[R3] Oh no! @BonjourMarc R U OK! please reply me for god's sake

[R5] Don't worry. I was home.

[R4] My gosh! that sucks:(Poor on u guys…

[R6] poor guys, terrible

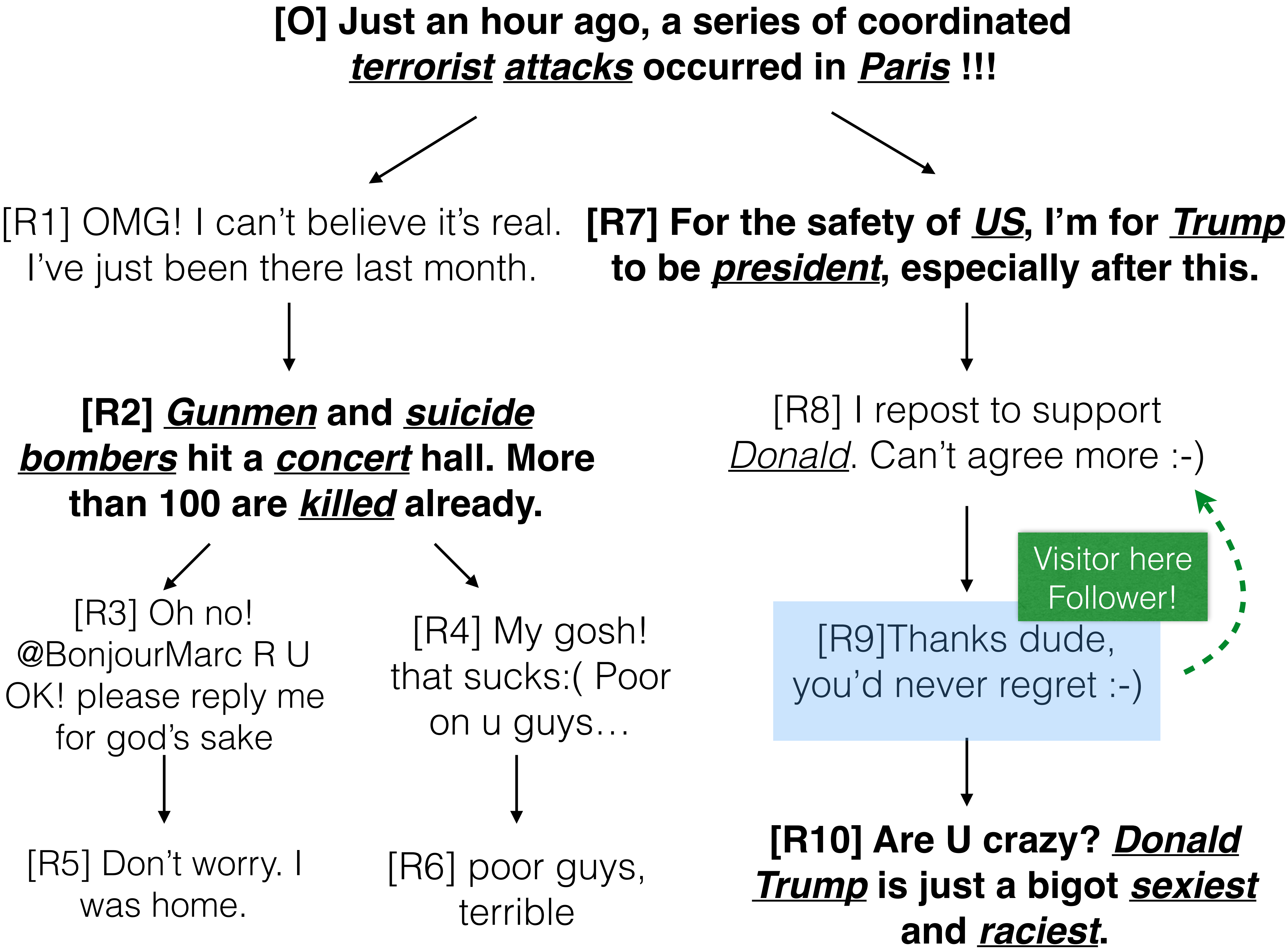
[R7] For the safety of *US*, I'm for *Trump* to be *president*, especially after this.

[R8] I repost to support *Donald*. Can't agree more :-)

Visitor here P_L=0.1

[R9]Thanks dude, you'd never regret :-)

[R10] Are U crazy? *Donald Trump* is just a bigot *sexiest* and *raciest*.



[O] Just an hour ago, a series of coordinated *terrorist attacks* occurred in *Paris* !!!

[R1] OMG! I can't believe it's real. I've just been there last month.

[R7] For the safety of *US*, I'm for *Trump* to be *president*, especially after this.

[R2] *Gunmen* and *suicide bombers* hit a *concert* hall. More than 100 are *killed* already.

[R8] I repost to support *Donald*. Can't agree more :-)

[R3] Oh no! @BonjourMarc R U OK! please reply me for god's sake

[R4] My gosh! that sucks:(Poor on u guys…

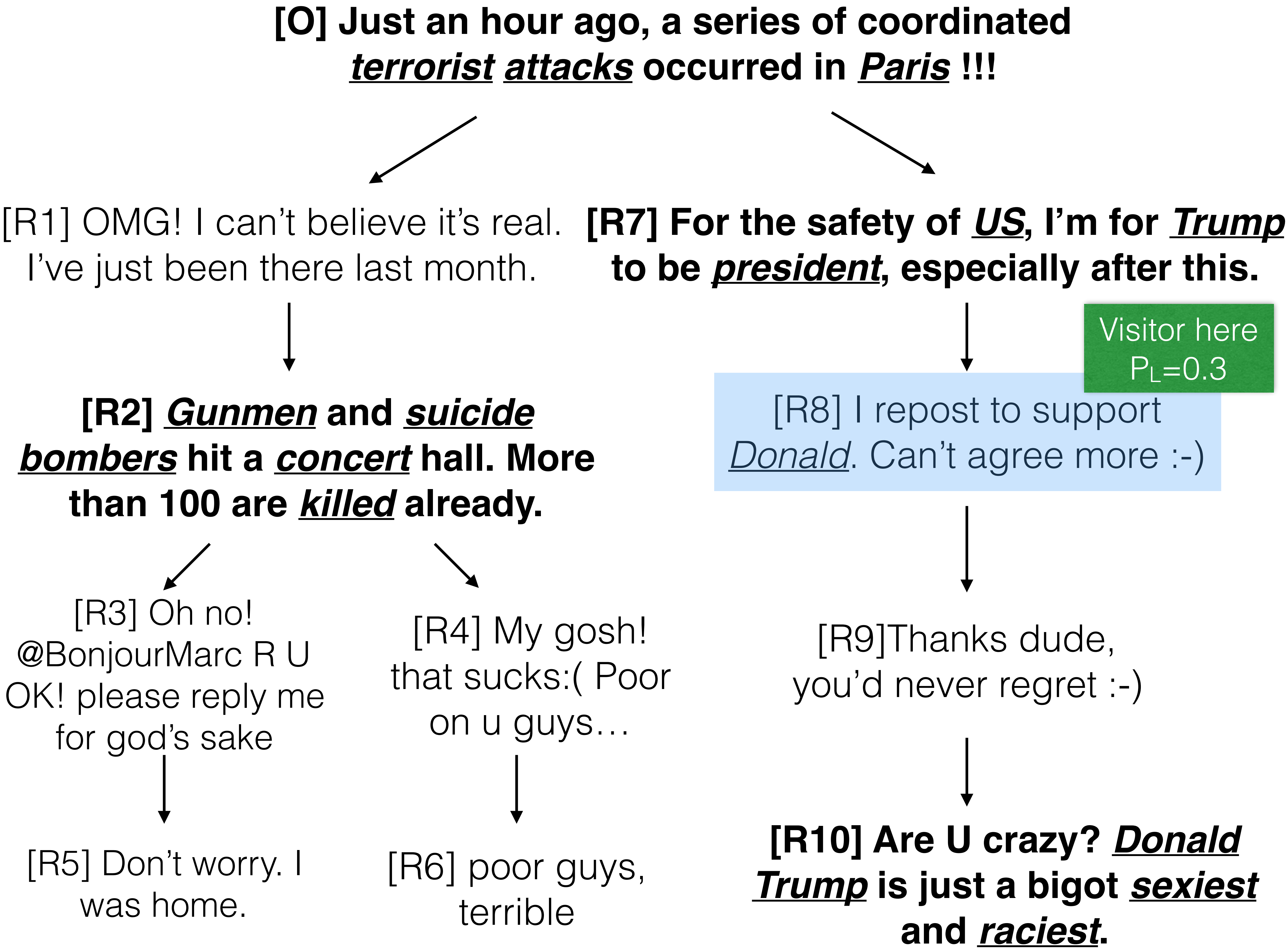
Visitor here Follower!

[R9]Thanks dude, you'd never regret :-)

[R5] Don't worry. I was home.

[R6] poor guys, terrible

[R10] Are U crazy? *Donald Trump* is just a bigot *sexiest* and *raciest*.



[O] Just an hour ago, a series of coordinated *terrorist attacks* occurred in *Paris* !!!

[R1] OMG! I can't believe it's real. I've just been there last month.

[R2] *Gunmen* and *suicide bombers* hit a *concert* hall. More than 100 are *killed* already.

[R3] Oh no! @BonjourMarc R U OK! please reply me for god's sake

[R5] Don't worry. I was home.

[R4] My gosh! that sucks:(Poor on u guys…

[R6] poor guys, terrible

[R7] For the safety of *US*, I'm for *Trump* to be *president*, especially after this.

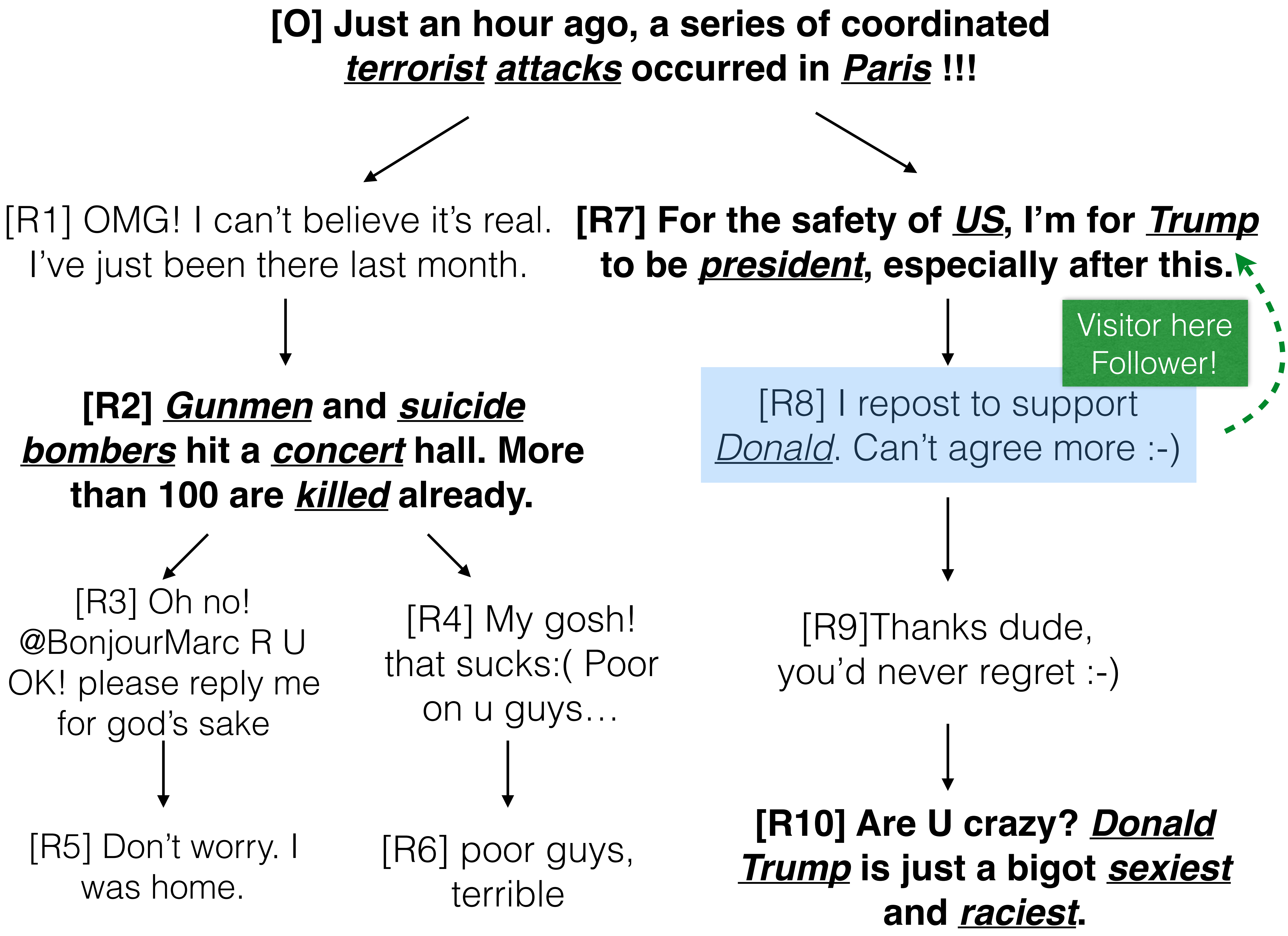
Visitor here

P_L=0.3

[R8] I repost to support *Donald*. Can't agree more :-)

[R9]Thanks dude, you'd never regret :-)

[R10] Are U crazy? *Donald Trump* is just a bigot *sexiest* and *raciest*.



[O] Just an hour ago, a series of coordinated *terrorist attacks* occurred in *Paris* !!!

[R1] OMG! I can't believe it's real. I've just been there last month.

[R2] *Gunmen* and *suicide bombers* hit a *concert* hall. More than 100 are *killed* already.

[R3] Oh no! @BonjourMarc R U OK! please reply me for god's sake

[R4] My gosh! that sucks:(Poor on u guys…

[R5] Don't worry. I was home.

[R6] poor guys, terrible

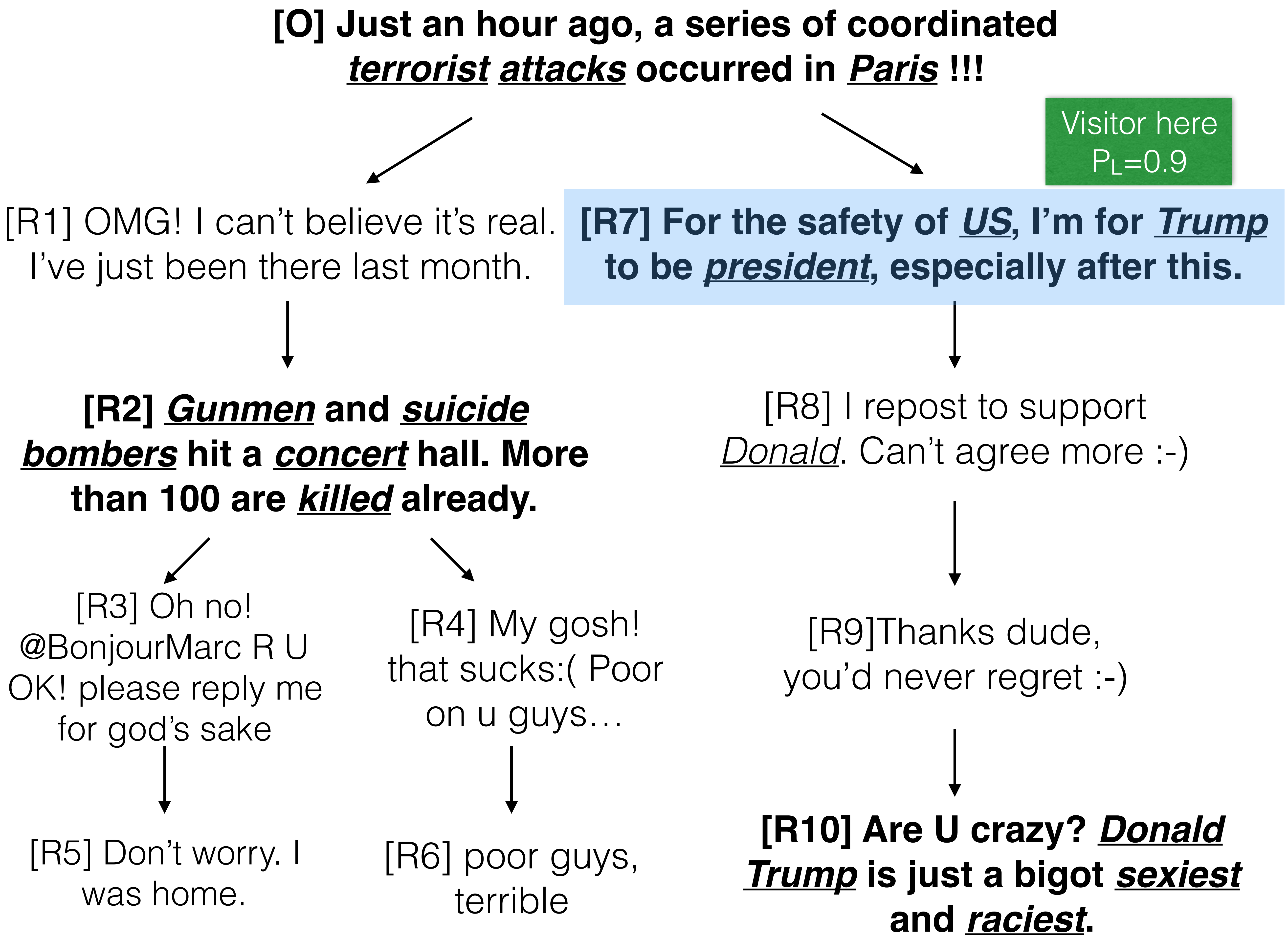
[R7] For the safety of *US*, I'm for *Trump* to be *president*, especially after this.

Visitor here
Follower!

[R8] I repost to support *Donald*. Can't agree more :-)

[R9]Thanks dude, you'd never regret :-)

[R10] Are U crazy? *Donald Trump* is just a bigot *sexiest* and *raciest*.



[O] Just an hour ago, a series of coordinated *terrorist attacks* occurred in *Paris* !!!

[R1] OMG! I can't believe it's real. I've just been there last month.

[R2] *Gunmen* and *suicide bombers* hit a *concert* hall. More than 100 are *killed* already.

[R3] Oh no! @BonjourMarc R U OK! please reply me for god's sake

[R5] Don't worry. I was home.

[R4] My gosh! that sucks:(Poor on u guys…

[R6] poor guys, terrible

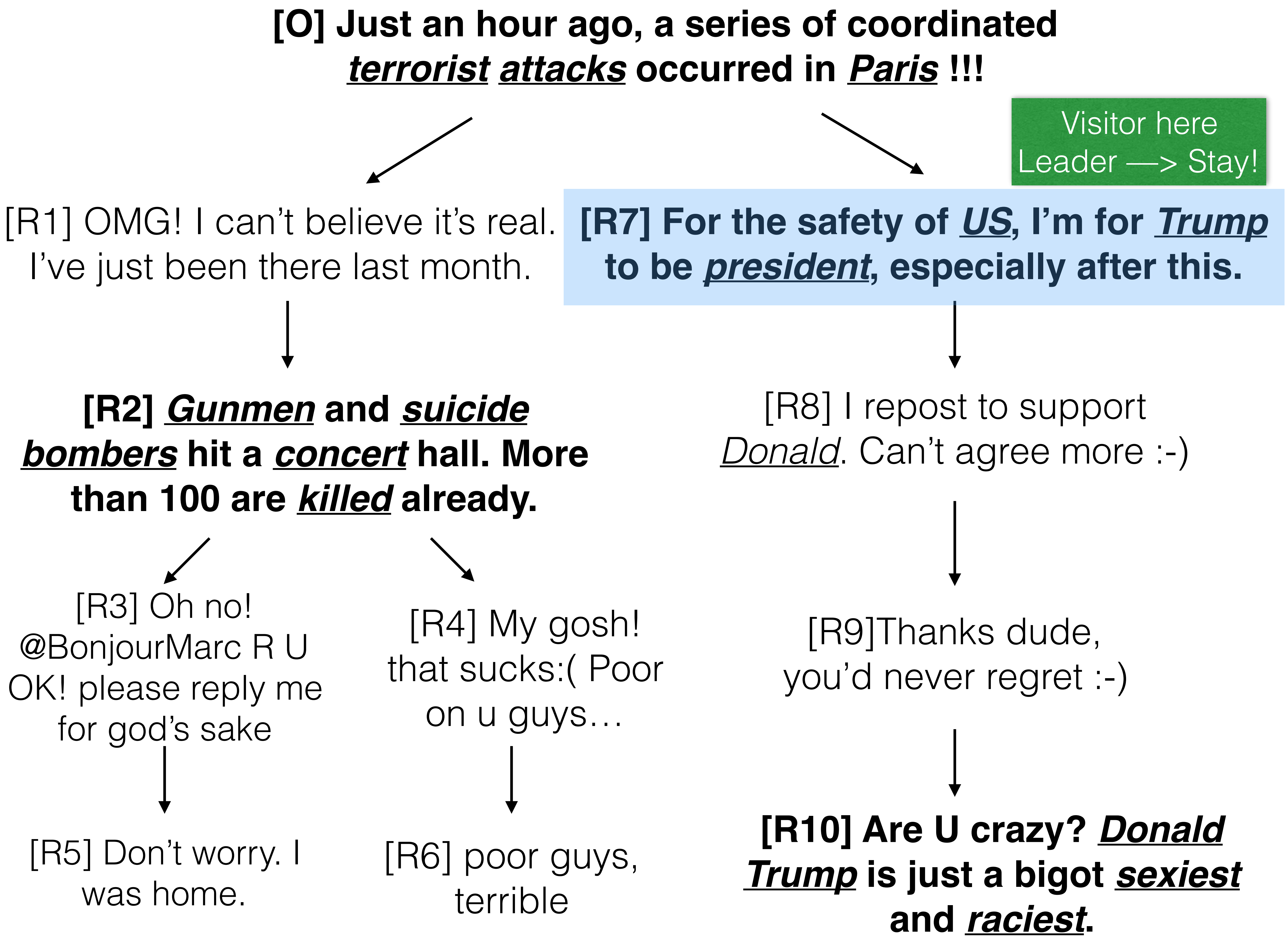
Visitor here P_L=0.9

[R7] For the safety of *US*, I'm for *Trump* to be *president*, especially after this.

[R8] I repost to support *Donald*. Can't agree more :-)

[R9]Thanks dude, you'd never regret :-)

[R10] Are U crazy? *Donald Trump* is just a bigot *sexiest* and *raciest*.



[O] Just an hour ago, a series of coordinated *terrorist attacks* occurred in *Paris* !!!

[R1] OMG! I can’t believe it’s real. I’ve just been there last month.

[R2] *Gunmen* and *suicide bombers* hit a *concert* hall. More than 100 are *killed* already.

[R3] Oh no! @BonjourMarc R U OK! please reply me for god’s sake

[R5] Don’t worry. I was home.

[R4] My gosh! that sucks:(Poor on u guys…

[R6] poor guys, terrible

Visitor here
Leader —> Stay!

[R7] For the safety of *US*, I’m for *Trump* to be *president*, especially after this.

[R8] I repost to support *Donald*. Can’t agree more :-)

[R9]Thanks dude, you’d never regret :-)

[R10] Are U crazy? *Donald Trump* is just a bigot *sexiest* and *raciest*.

Outline

- **Conversation Tree Summarization**
  - Introduction
  - LeadSum Summarization Model
  - **Experiments**

Data Collections for Summarization

Name	# of nodes	# of nodes with	Height	Category
Tree (I)	21,353	15,409	16	Social News
Tree (II)	9,616	6,073	11	Social News
Tree (III)	13,087	9,583	8	Movie
Tree (IV)	12,865	7,083	8	Music
Tree (V)	10,666	7,129	8	Entertainment news
Tree (VI)	21,127	15,057	11	Sports news
Tree (VII)	18,974	12,399	13	Social news
Tree (VIII)	2,021	925	18	Political news
Tree (IX)	9,230	5,408	14	Breaking event
Tree (X)	10,052	4,257	25	Breaking event

Evaluation on Summarization

	ROUGE-1	ROUGE-2
RandSum	0.159	0.037
RepRankSum	0.162	0.030
UserRankSum	0.292	0.087
LeadProSum	0.270	0.064
SVDSum	0.222	0.048
DivRankSum	0.159	0.029
GBDTSum	0.272	0.071
Basic-LeadSum	0.300	0.082
Soft-LeadSum	**0.351**	**0.105**

HIV research	This news is terribly shocking. Losing these AIDS experts would be a great loss for all human-beings.
	This crash brings another big blow to HIV research after the event of ``Mississippi baby". This is a tragedy for the whole human society. Let's prey for all the victims.
Scientists	Those guys who would bring great contribution to medical research turned out to become victims of wars. I feel grieved for that.
Conjecture	Some of these biologists may have developed some dangerous Gene medicine by chance. Future guys know about this and travel through time to stop the production of this Gene medicine using this crash.
Background	There are 108 HIV experts, researchers and their family killed in this crash. They prepare to land in Kuala Lumpur and transfer to Melbourne to attend the 20th AIDS conference. Organizing committee of AIDS released letter of condolence.
Life	All men are created equal. But some people may contribute a lot more to our world.
Suggestion	I think top experts should not be allowed to take the same plane all together.
Malaysia Airlines	There are many excellent artists on MH370. I feel that Malaysia Airlines may have some conspiracy.
Opinion	This makes things even worse than a crash.
War	Great loss to human-beings. No war is good. It only brings disaster.

Outline

- Background
- Microblog Topic Extraction
- Conversation Tree Summarization
- **Conclusion**

Conclusion

- Summarization framework based on conversation structures to enrich contextual information.

- Conversation modeling: differentiate posts as leaders and followers in context of conversation trees

- A novel topic model considering conversation structures, which benefits down-stream applications.

- Datasets for leader detection, summarization and topic modeling on microblog conversations.

Reference

- Jing Li, Ming Liao, Wei Gao, Yulan He, Kam-Fai Wong: Topic Extraction from Microblog Posts Using Conversation Structures. ACL (1) 2016

- Jing Li, Wei Gao, Zhongyu Wei, Baolin Peng, Kam-Fai Wong: Using Content-level Structures for Summarizing Microblog Repost Trees. EMNLP 2015: 2168-2178

Reference

- (Yan et al. 13) Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng: A biterm topic model for short texts. WWW 2013: 1445-1456

- (Quan et al. 15) Xiaojun Quan, Chunyu Kit, Yong Ge, Sinno Jialin Pan: Short and Sparse Text Topic Modeling via Self-Aggregation. IJCAI 2015: 2270-2276

Reference

- (Chang et al. 13) Yi Chang, Xuanhui Wang, Qiaozhu Mei, Yan Liu: Towards Twitter context summarization with user influence models. WSDM 2013: 527-536

- (Lafferty et al. 01) John D. Lafferty, Andrew McCallum, Fernando C. N. Pereira: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML 2001: 282-289

- (Mei et al. 10) Qiaozhu Mei, Jian Guo, Dragomir R. Radev: DivRank: the interplay of prestige and diversity in information networks. KDD 2010: 1009-1018

Thanks
谢谢

lijing@se.cuhk.edu.hk
http://www1.se.cuhk.edu.hk/~lijing/