Decentralized Entity-Level Modeling for Coreference Resolution

Greg Durrett, David Hall, and Dan Klein

UC Berkeley

Entity-Level Modeling

Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

[his]

Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

James Reed	Rose Brooks	Reed	Brooks	his

Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

MALE	FEMALE			MALE
James Reed	Rose Brooks	Reed	Brooks	his

Entity-Level Modeling

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

Pairwise Approach

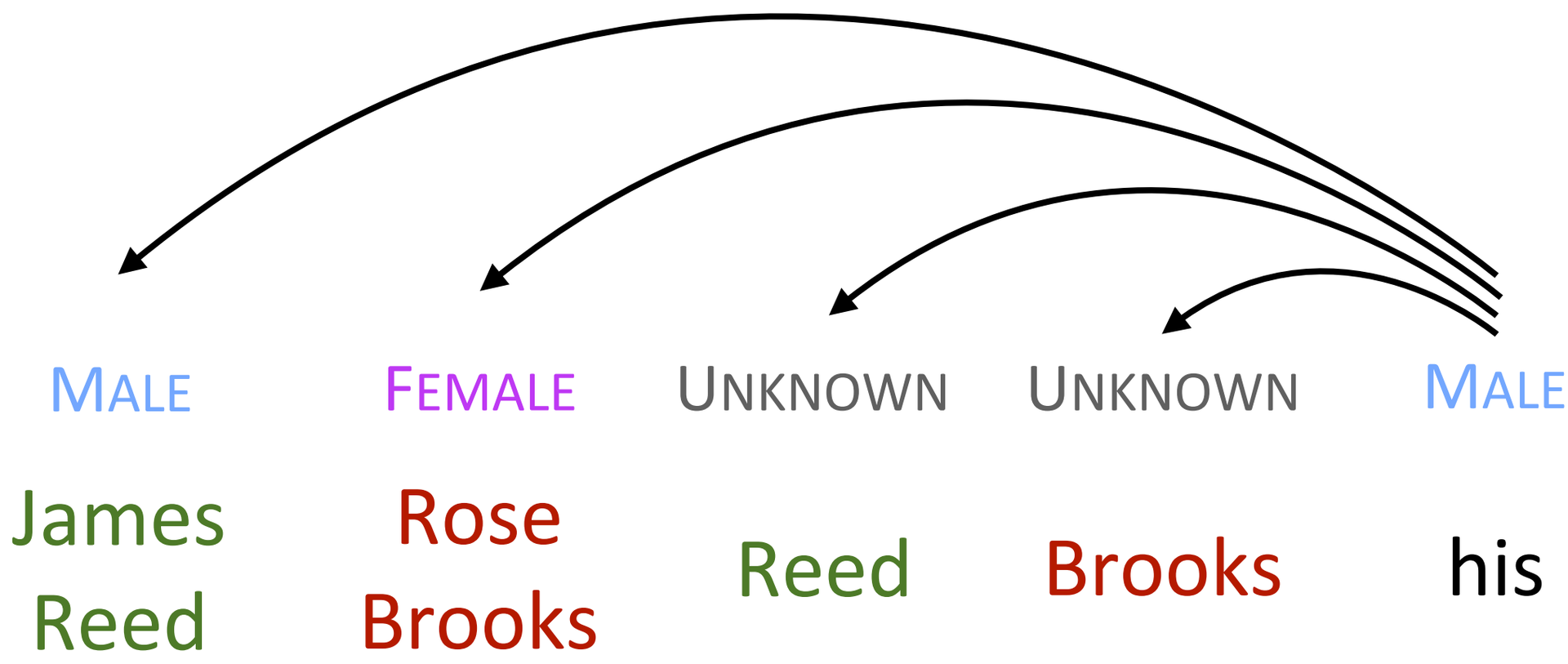
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Soon et al. (2001) inter alia]

Pairwise Approach

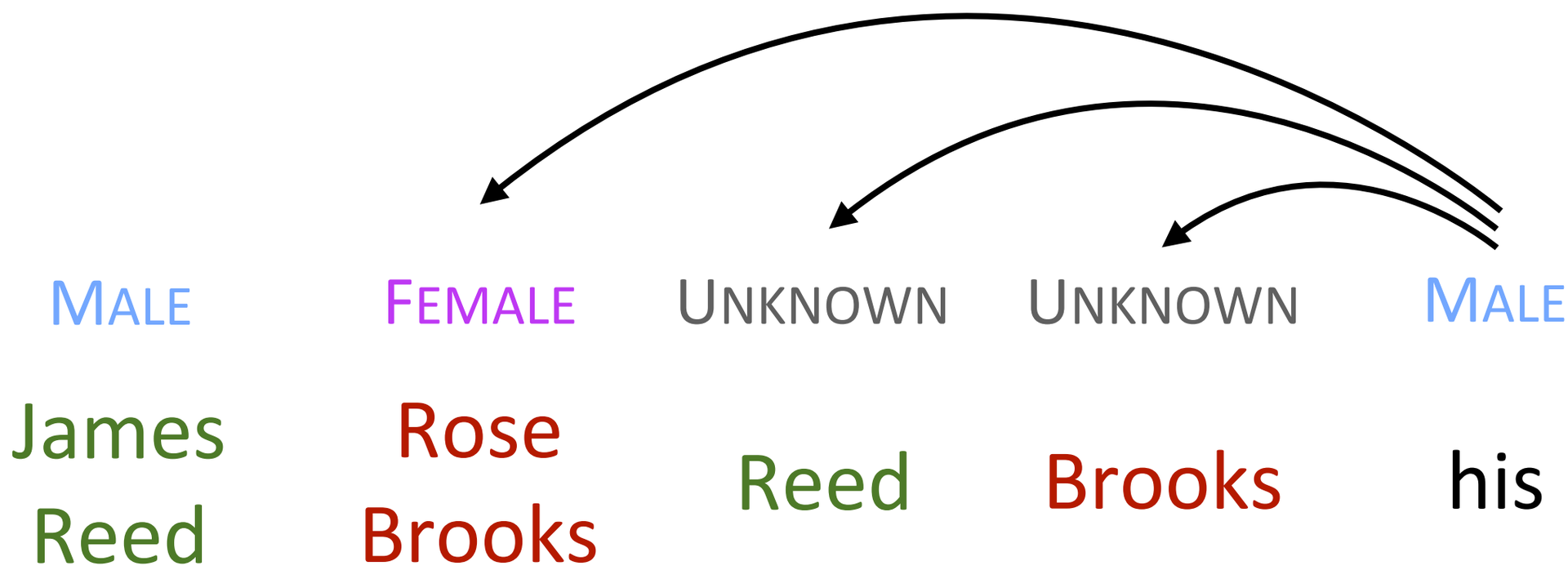
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



[Soon et al. (2001) inter alia]

Pairwise Approach

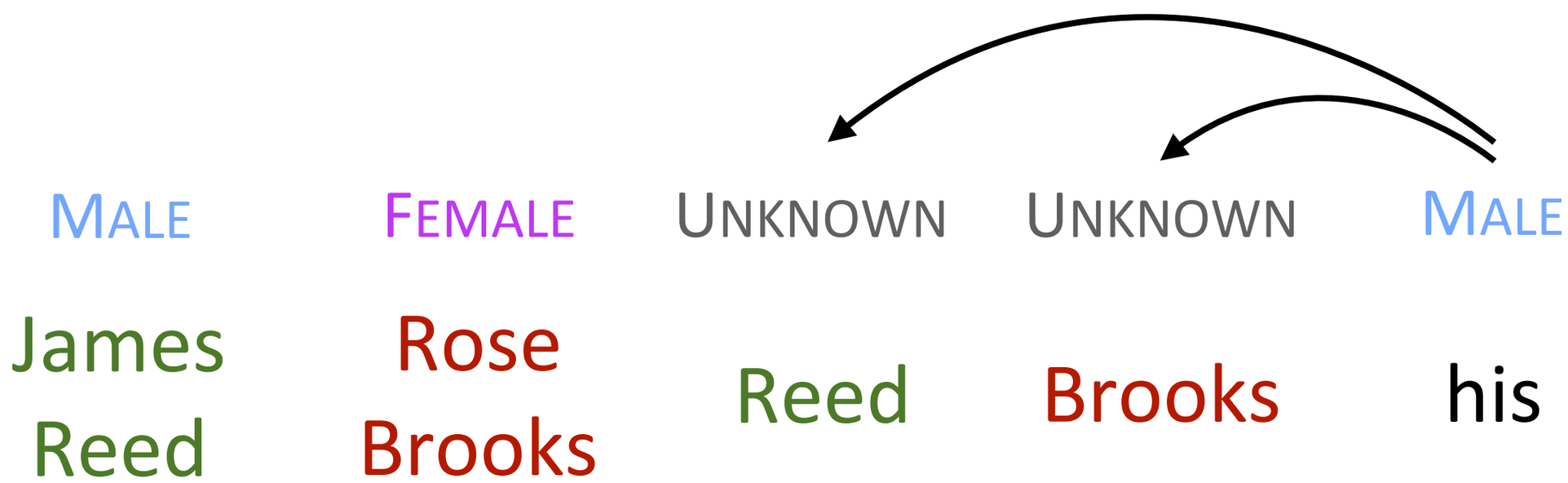
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



[Soon et al. (2001) inter alia]

Pairwise Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



[Soon et al. (2001) inter alia]

Pairwise Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Soon et al. (2001) inter alia]

Pairwise Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Does not propagate information

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Soon et al. (2001) inter alia]

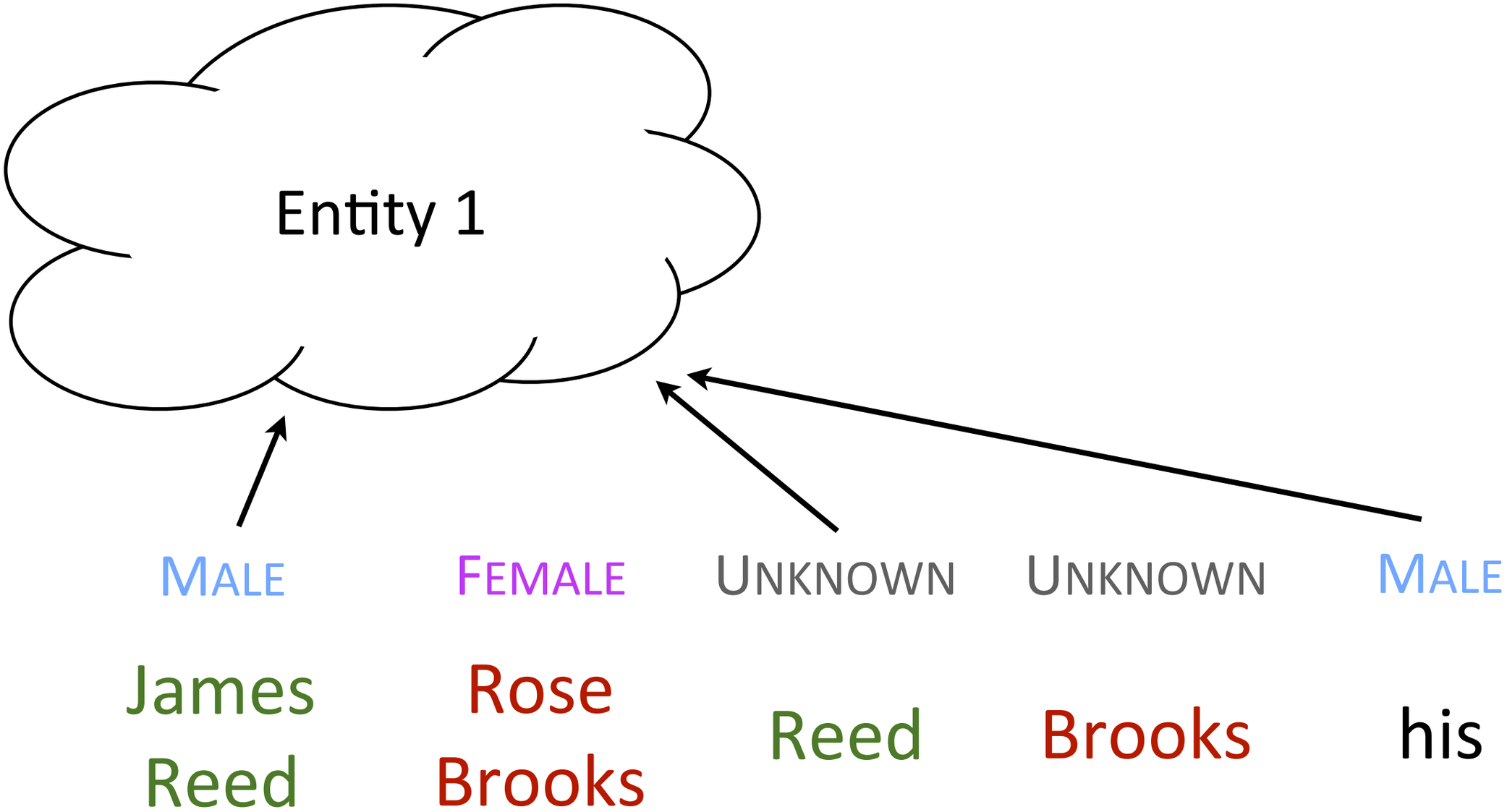
Entity-Level Models

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

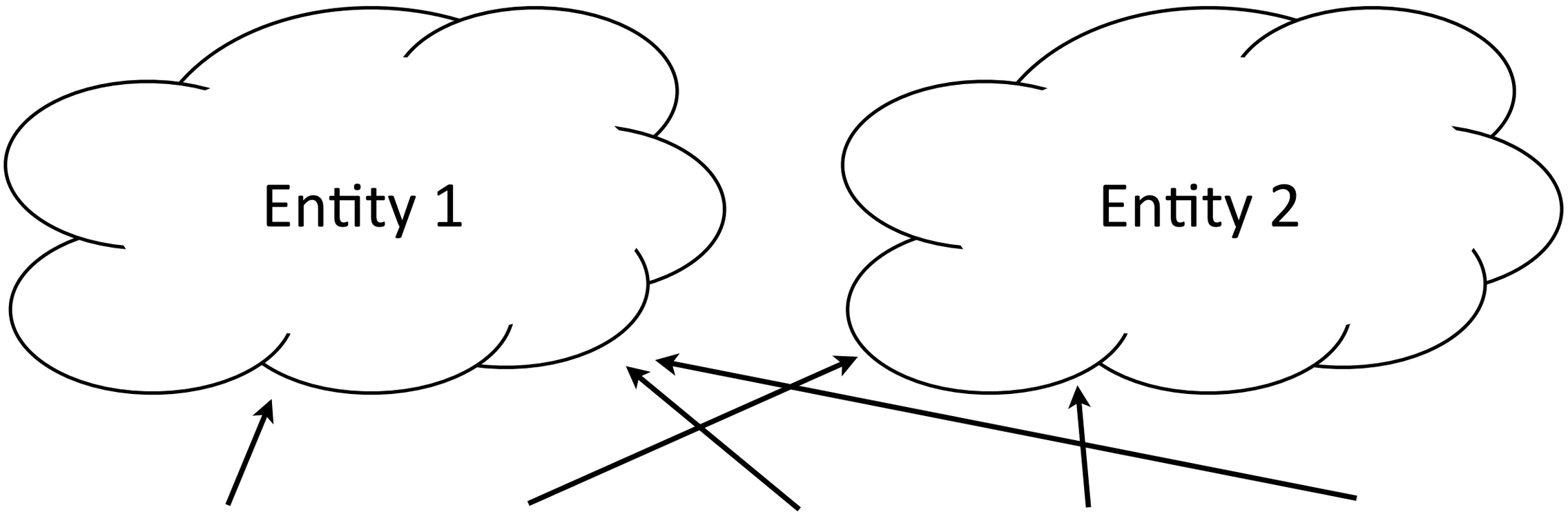
Entity-Level Models

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



Entity-Level Models

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

Centralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Luo et al. (2004), Rahman and Ng (2009)]

Centralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Gender: MALE
[James Reed]

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Luo et al. (2004), Rahman and Ng (2009)]

Centralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Gender: MALE
[James Reed]

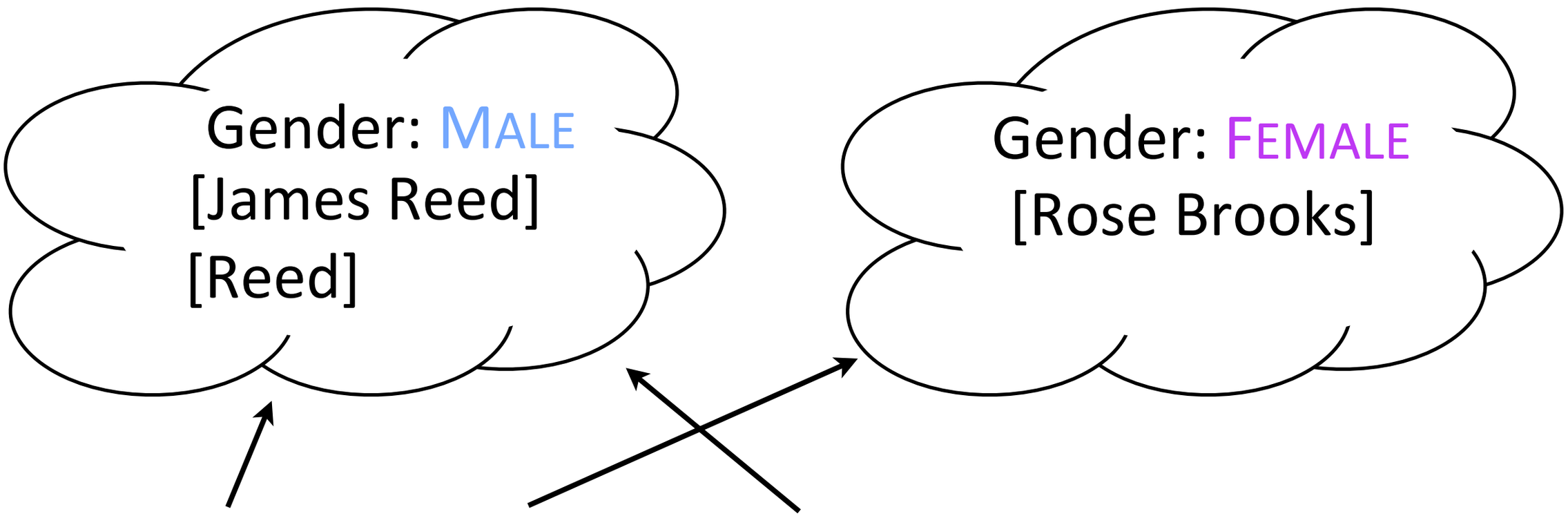
Gender: FEMALE
[Rose Brooks]

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Luo et al. (2004), Rahman and Ng (2009)]

Centralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

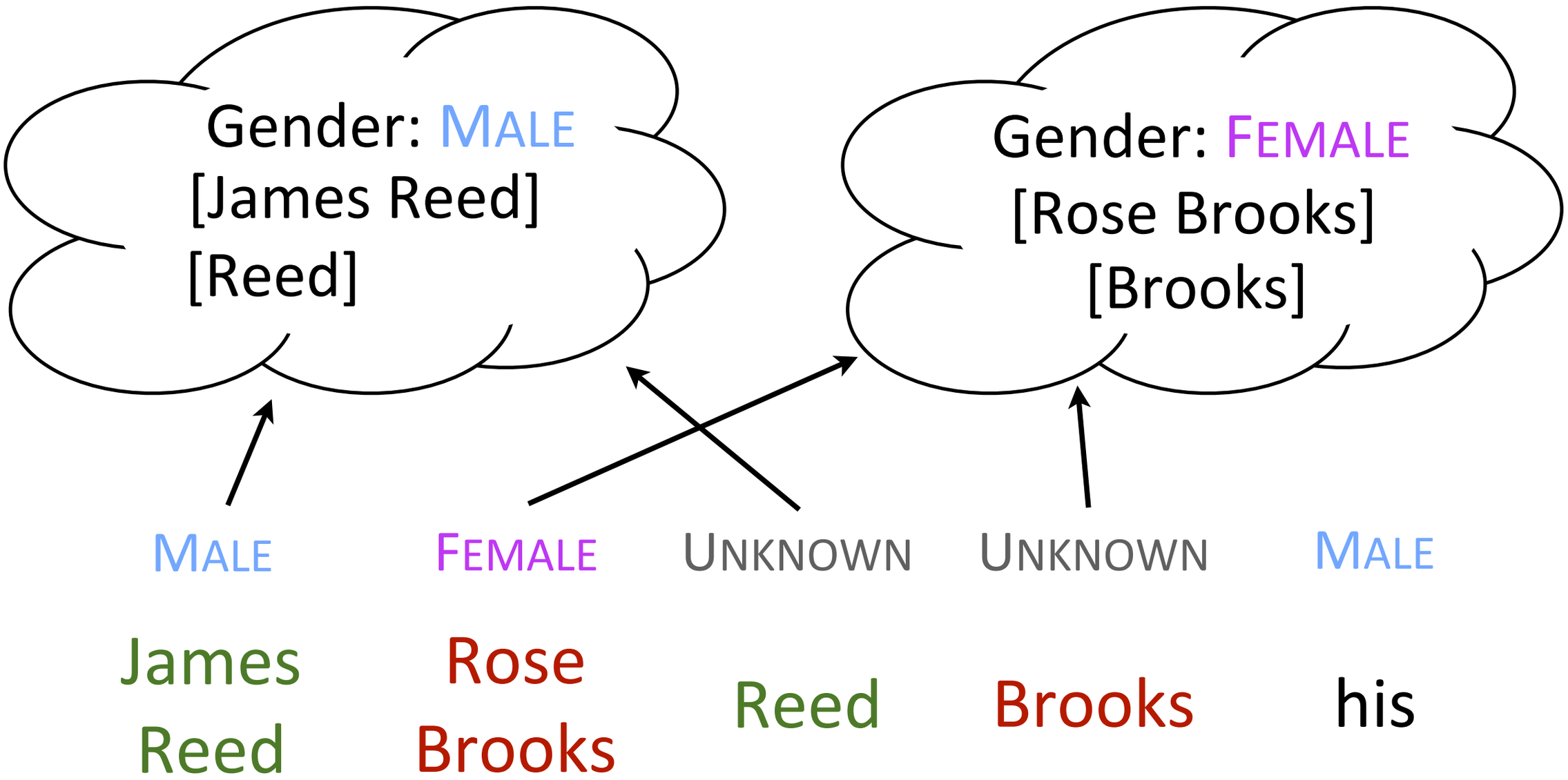


MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Luo et al. (2004), Rahman and Ng (2009)]

Centralized Approach

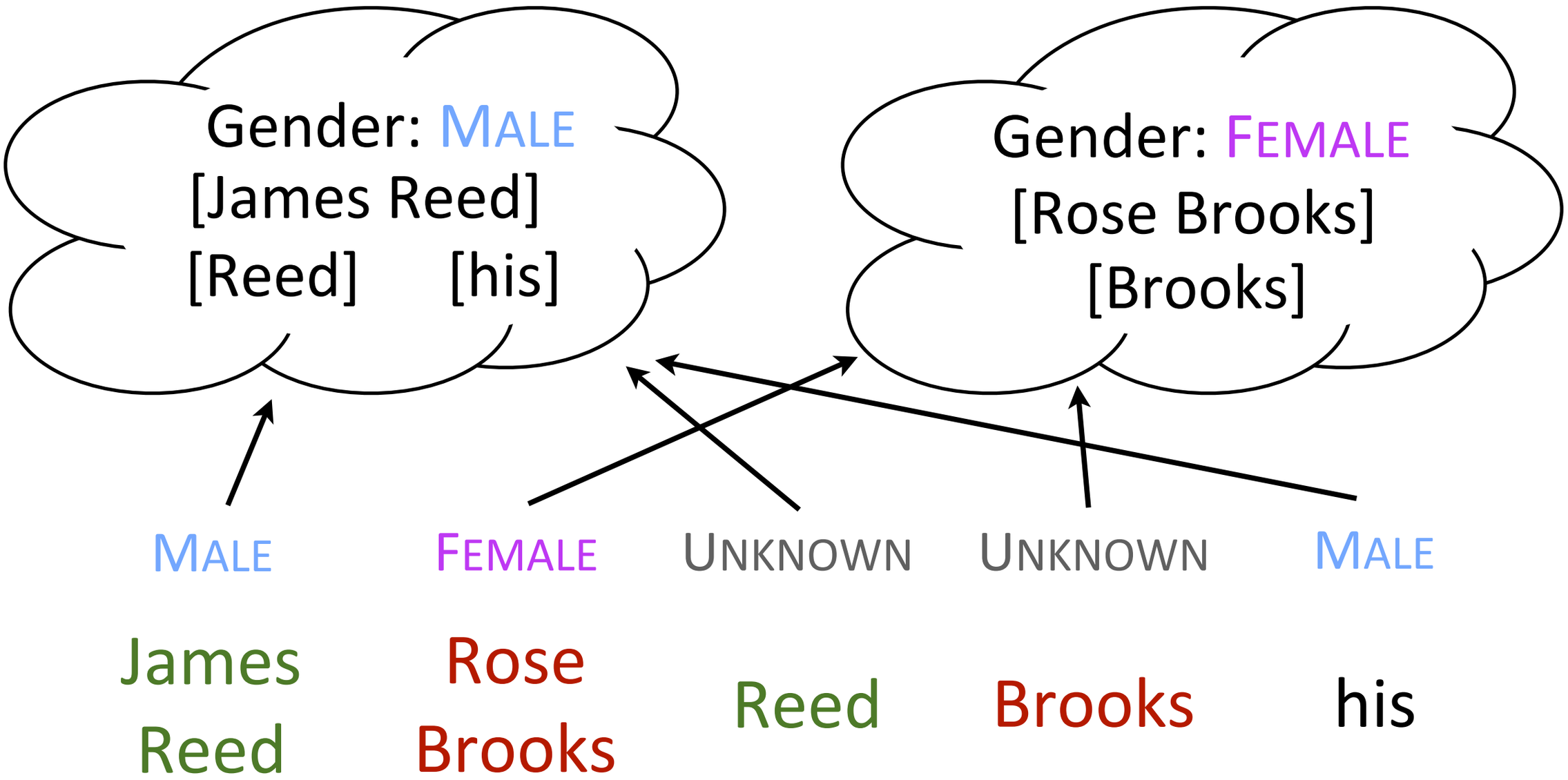
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



[Luo et al. (2004), Rahman and Ng (2009)]

Centralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



[Luo et al. (2004), Rahman and Ng (2009)]

Centralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Gender: MALE
[James
[Reed]

Gender: FEMALE
s]

Does not maintain uncertainty during inference

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

[Luo et al. (2004), Rahman and Ng (2009)]

Our Decentralized Approach

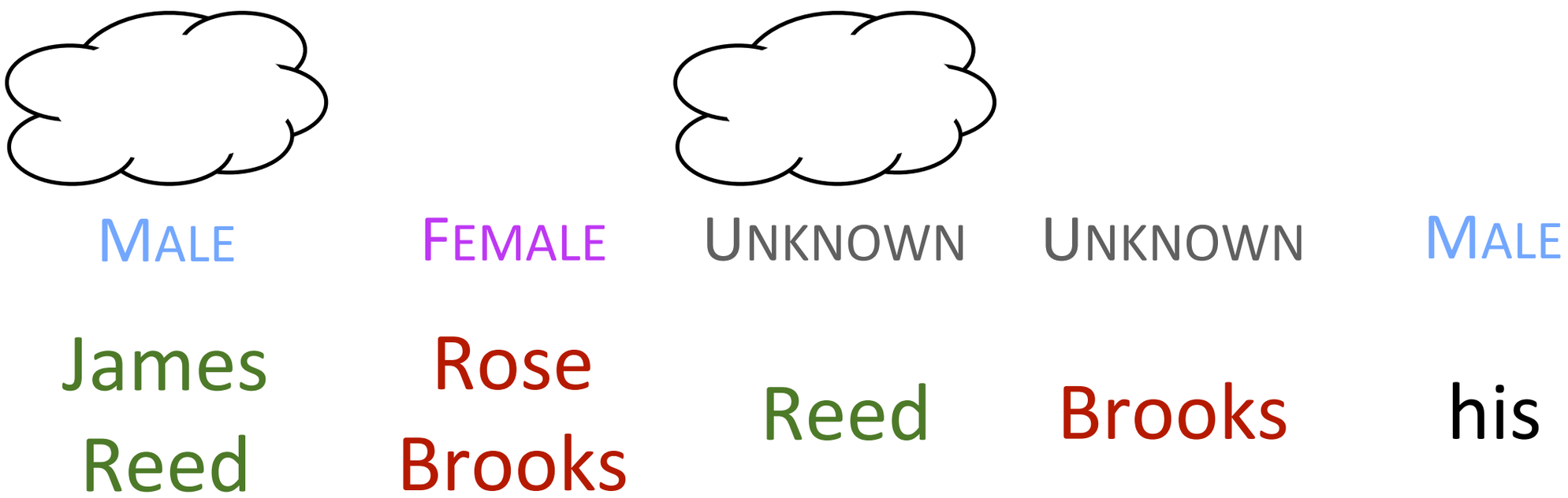
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Gender: MALE

MALE	FEMALE	UNKNOWN	UNKNOWN	MALE
James Reed	Rose Brooks	Reed	Brooks	his

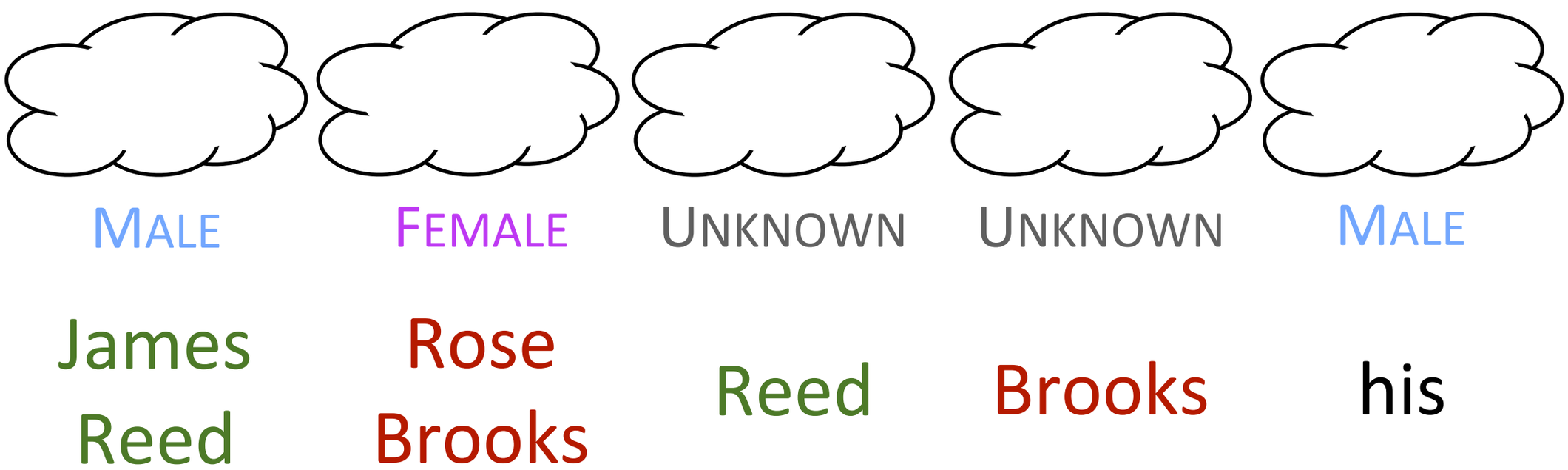
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



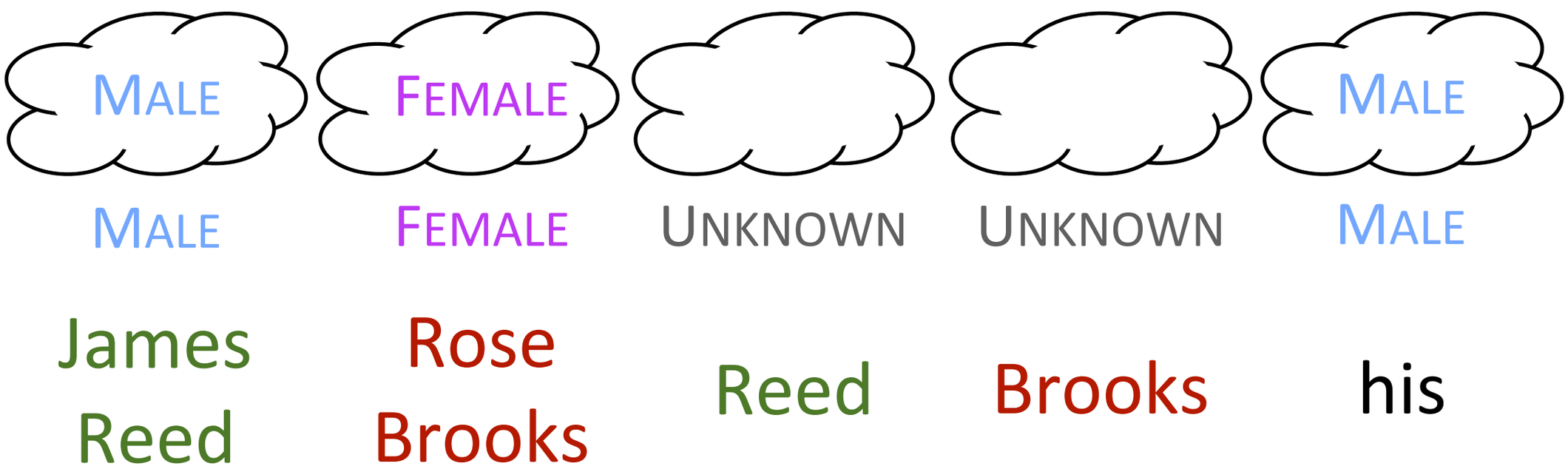
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



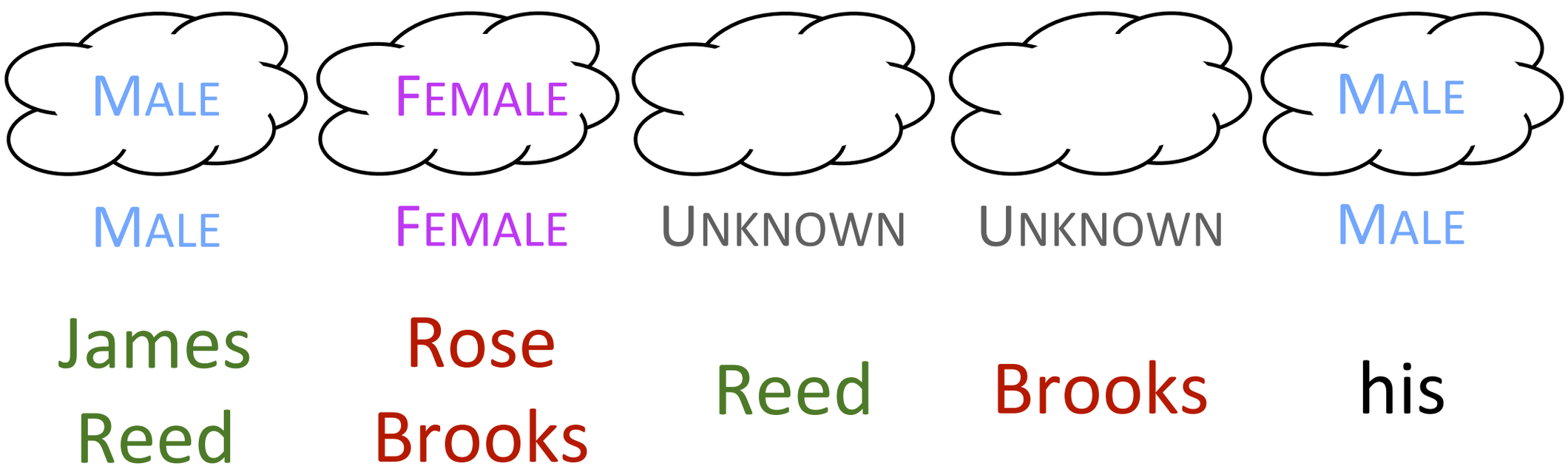
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



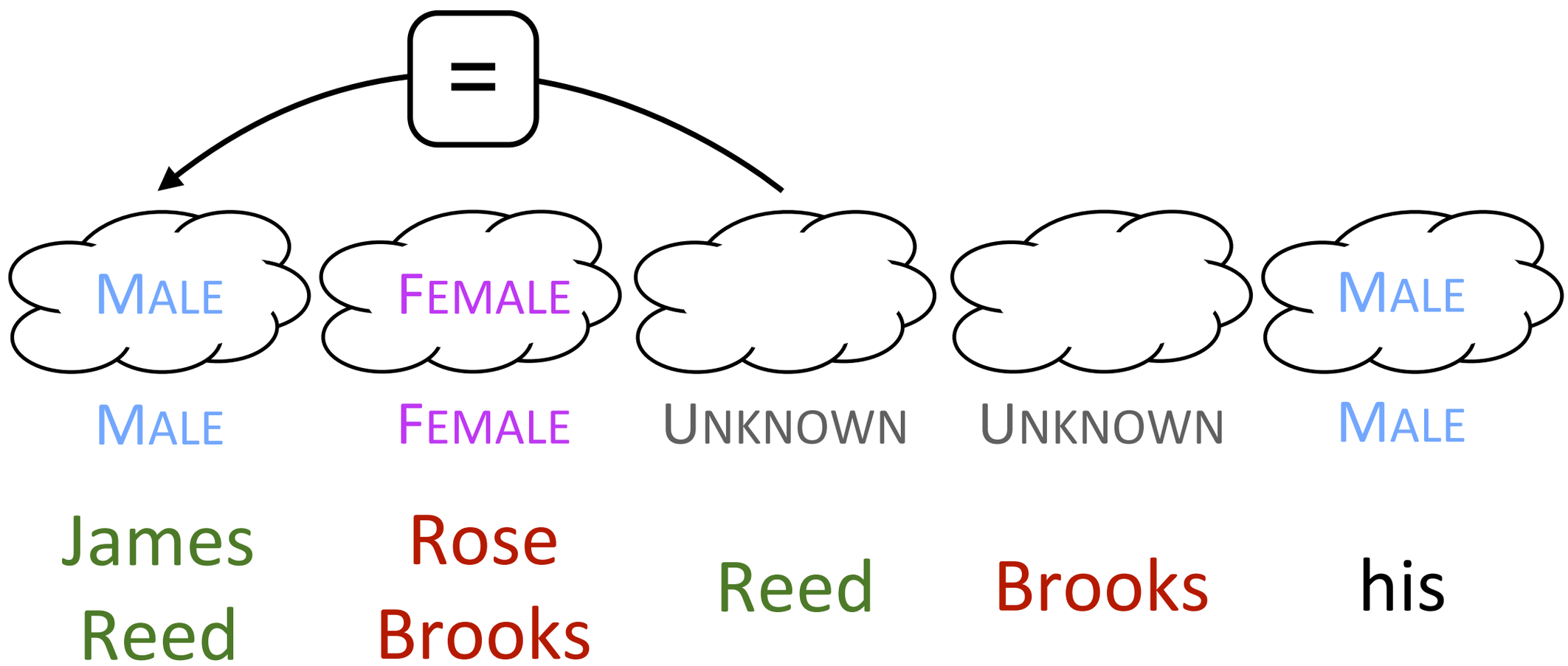
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



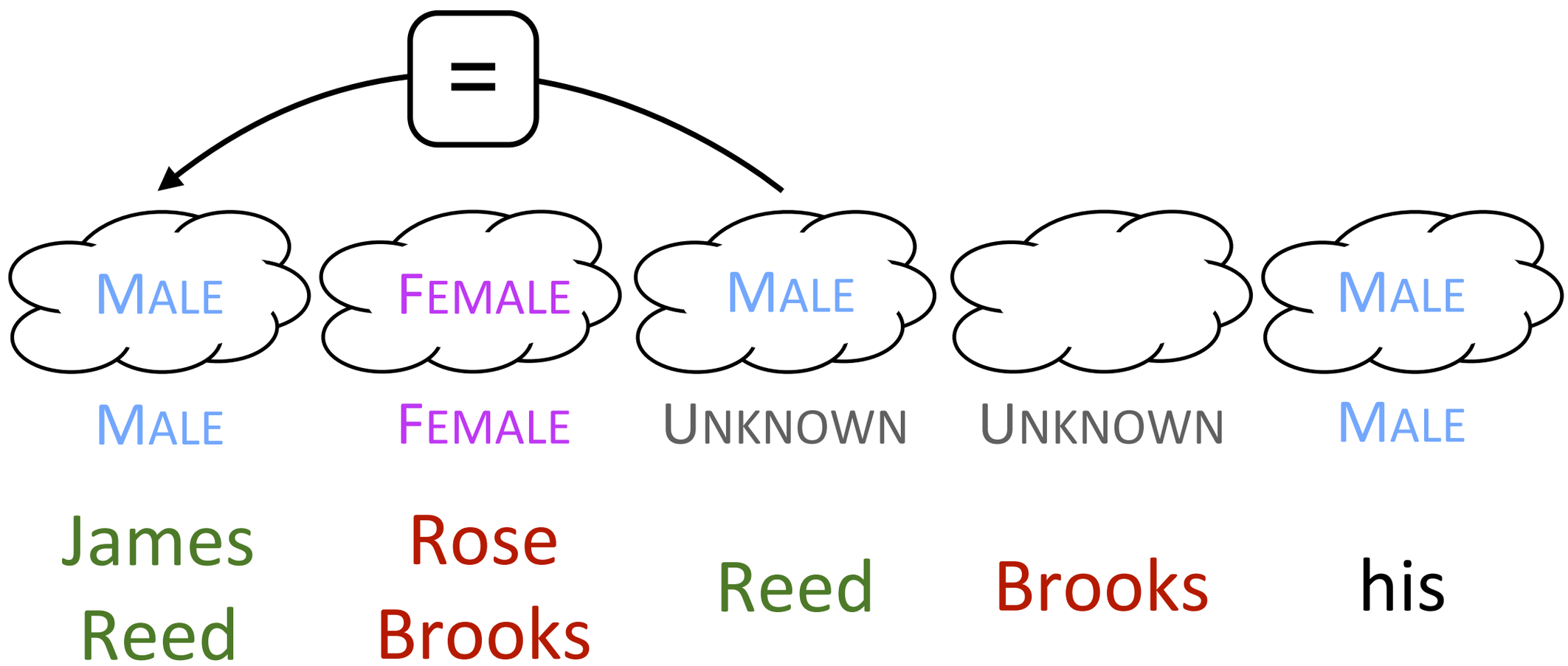
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



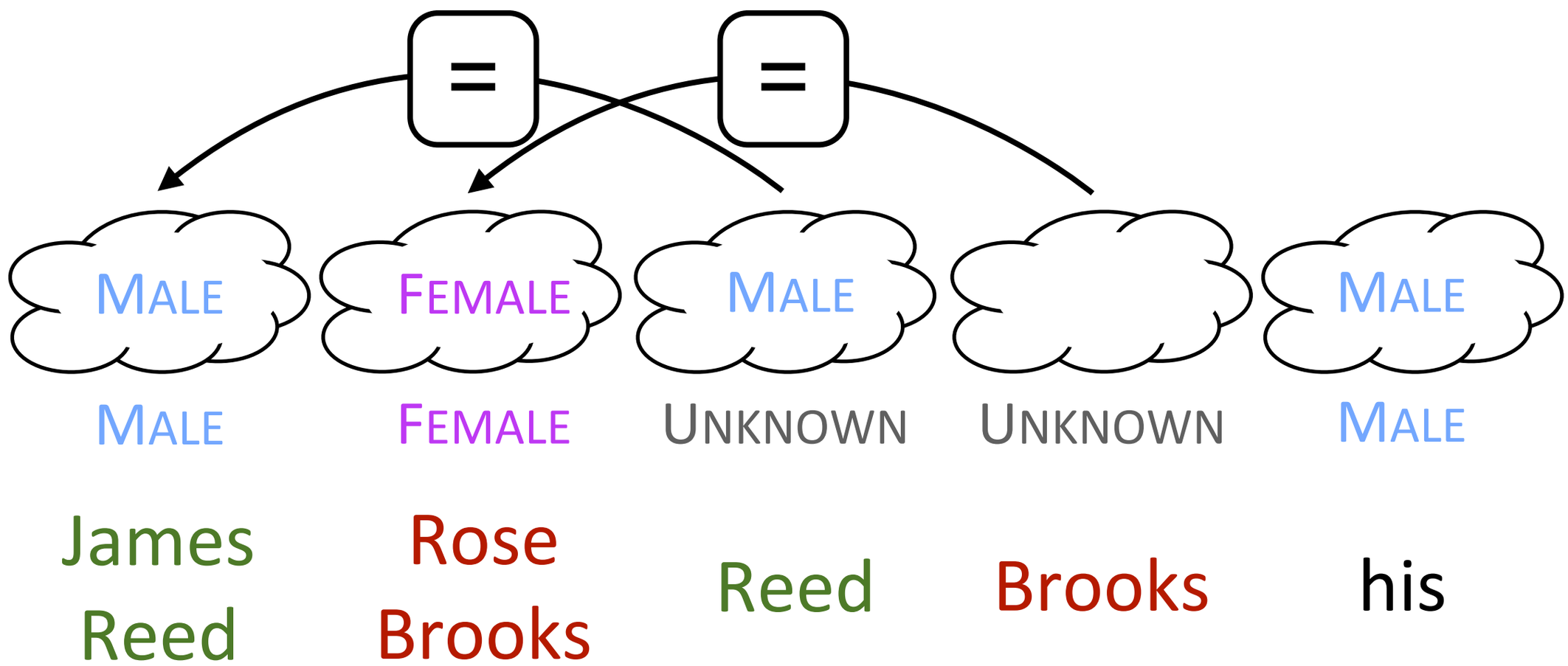
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



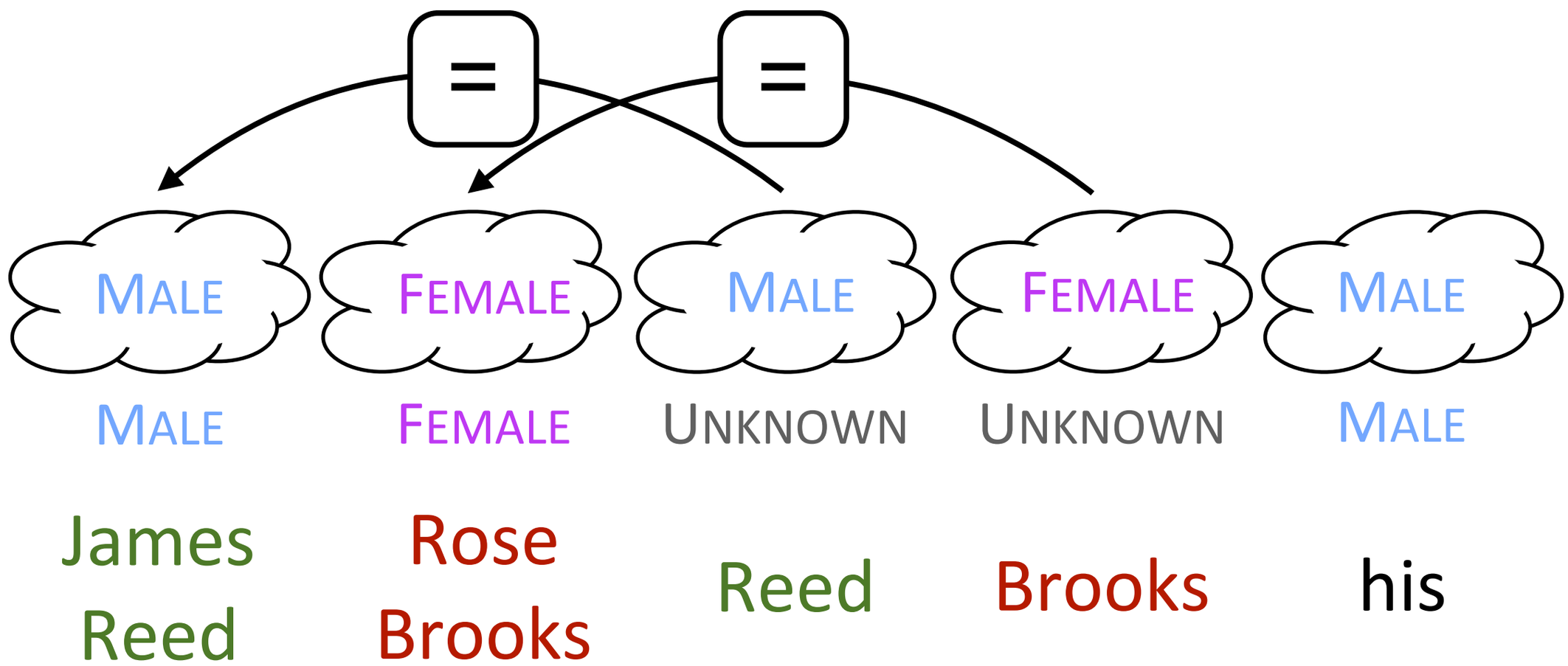
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



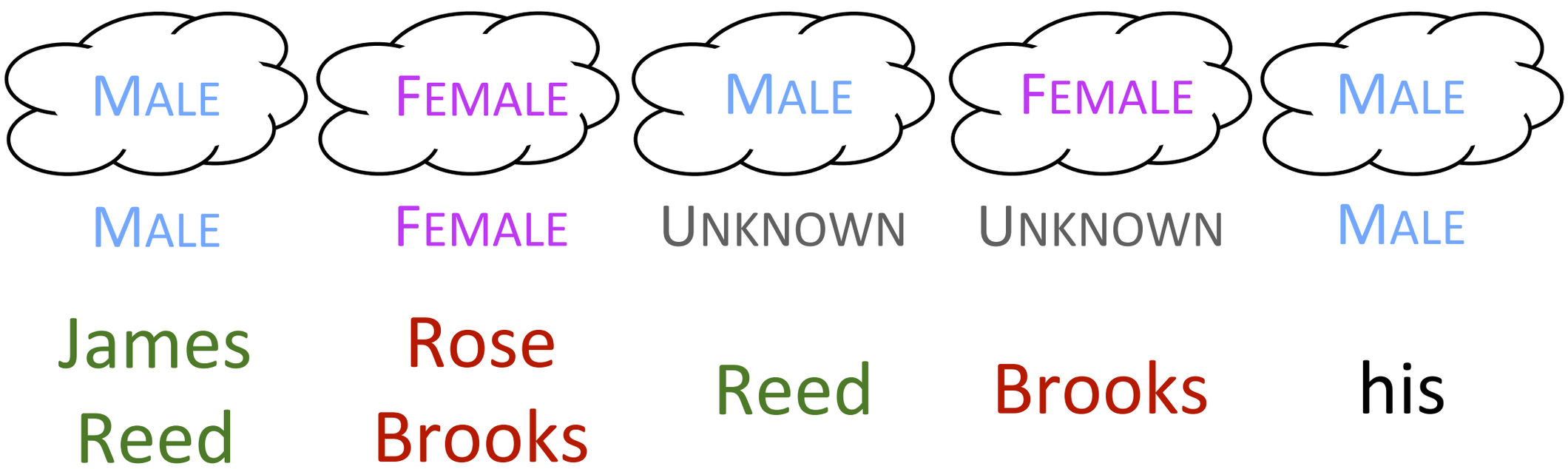
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



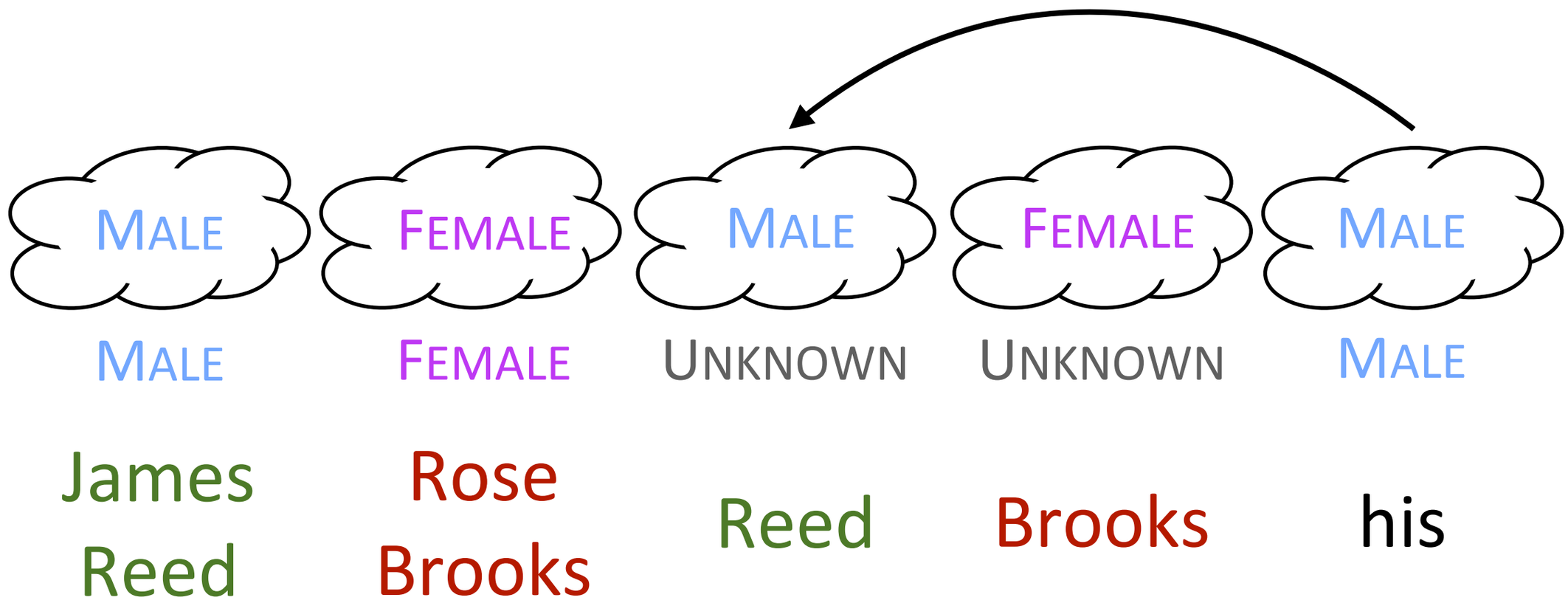
Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



Our Decentralized Approach

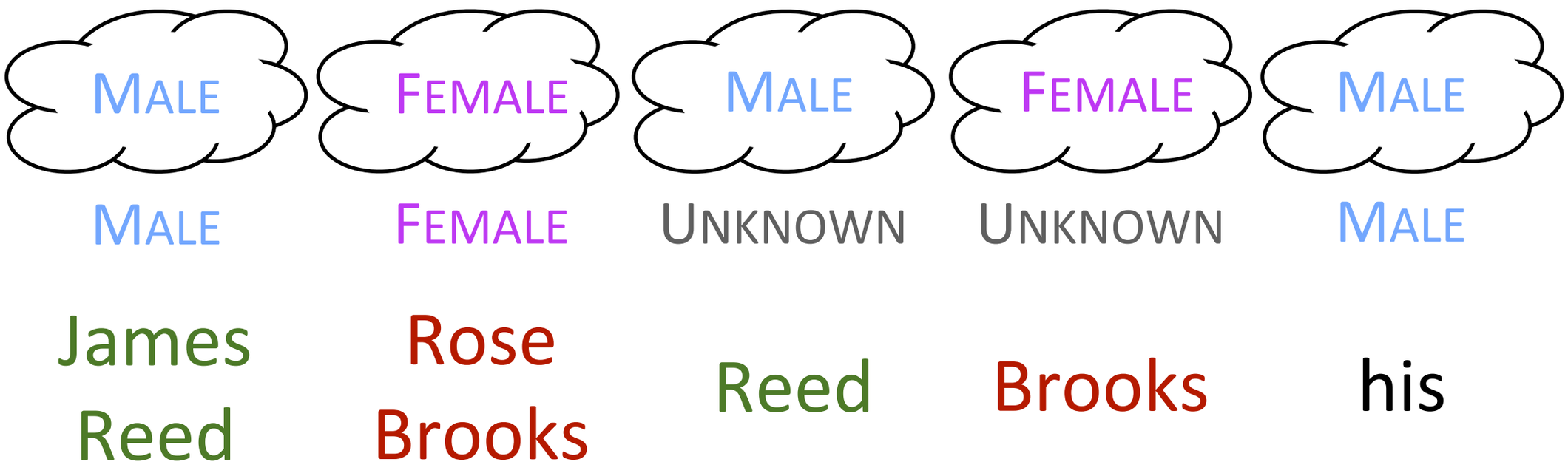
New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.



Our Decentralized Approach

New York was where [James Reed] met [Rose Brooks]. [Reed] was introduced to [Brooks] at [his] company's Christmas party.

Maintains tractability of pairwise system, incorporates entity-level information



Basic Model

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

[Denis and Baldridge (2008)]

Basic Model

New

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

[Denis and Baldridge (2008)]

Basic Model

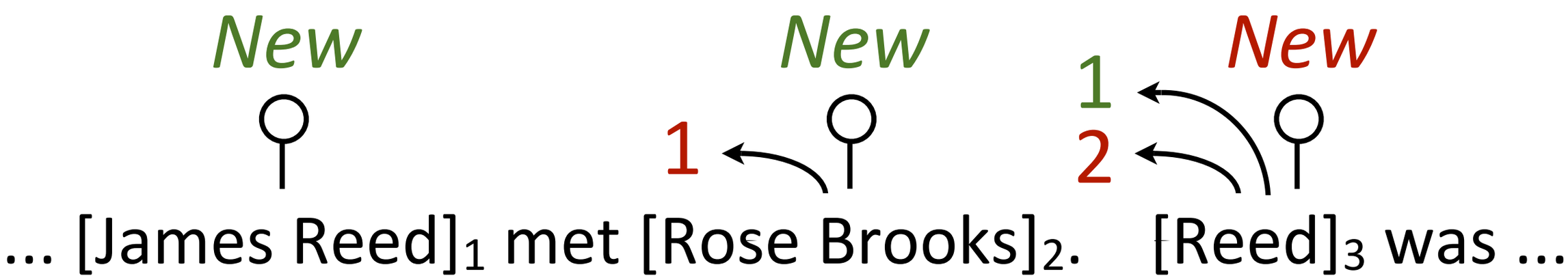
New *New*

1

... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

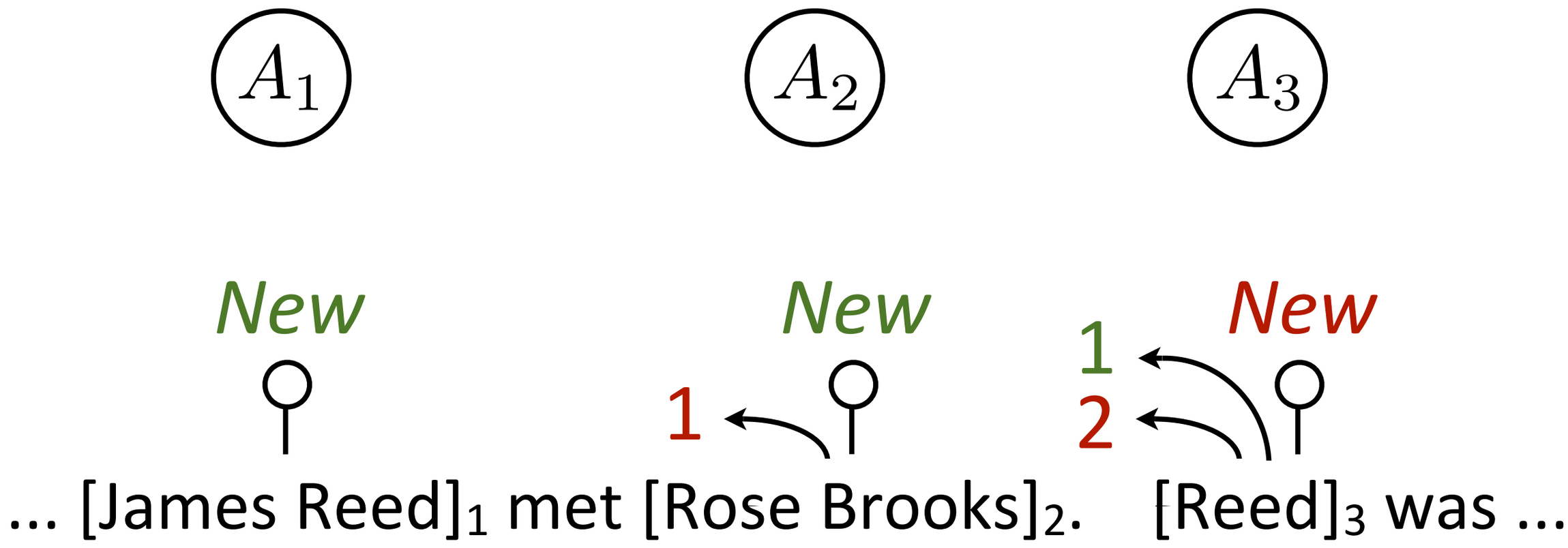
[Denis and Baldridge (2008)]

Basic Model



[Denis and Baldridge (2008)]

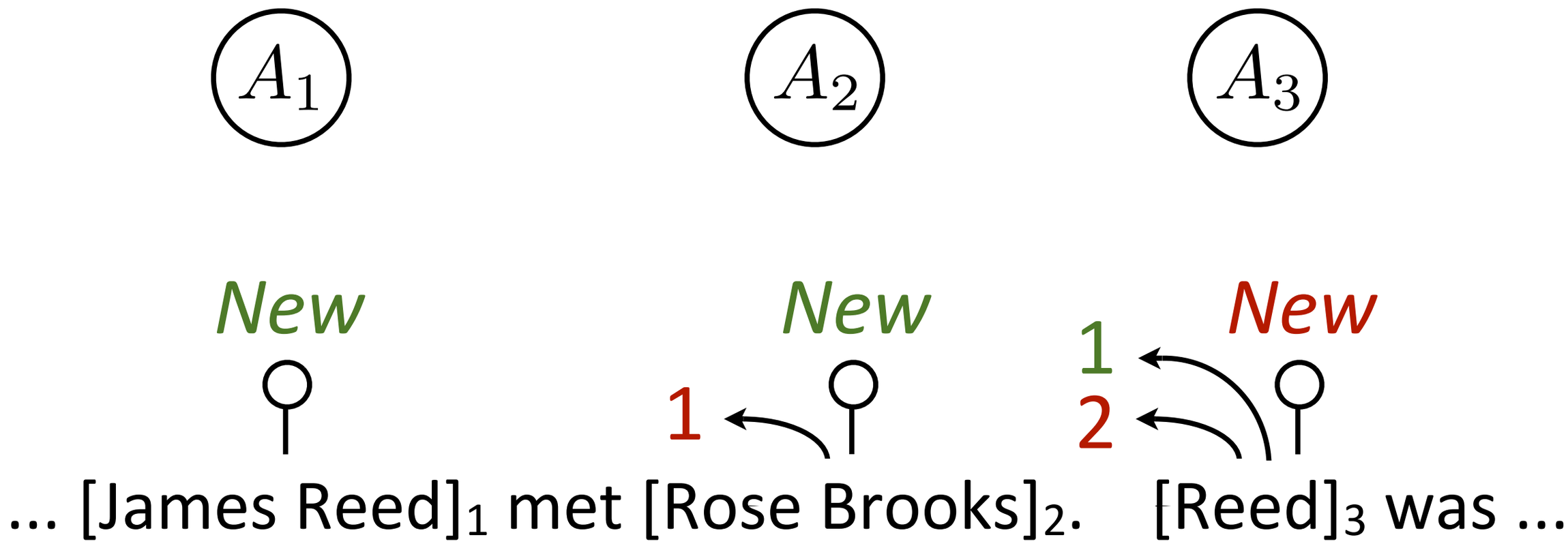
Basic Model



[Denis and Baldridge (2008)]

Basic Model

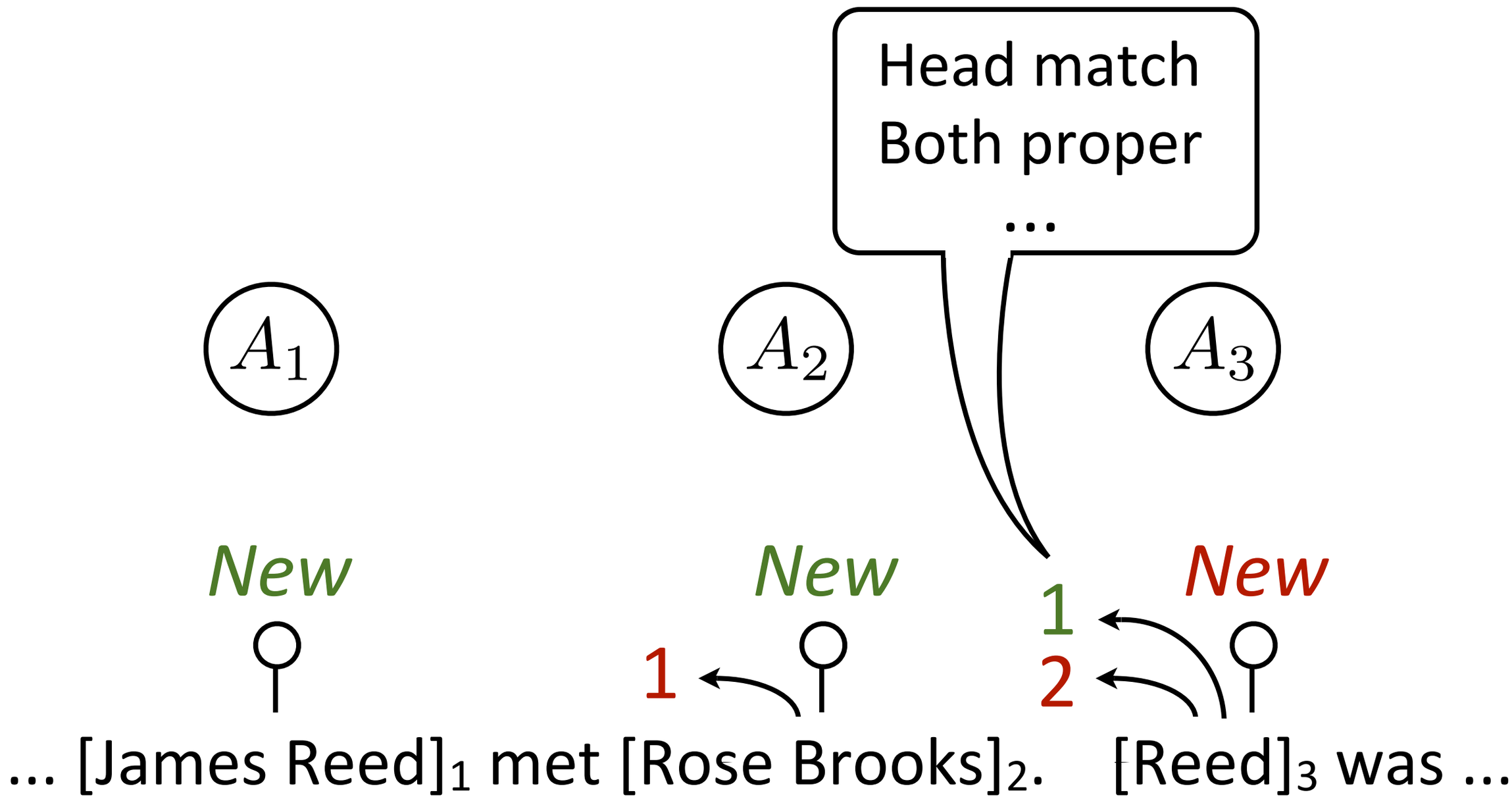
$$Pr(a_i|x) \propto \exp(w^T f(a_i, x))$$



[Denis and Baldridge (2008)]

Basic Model

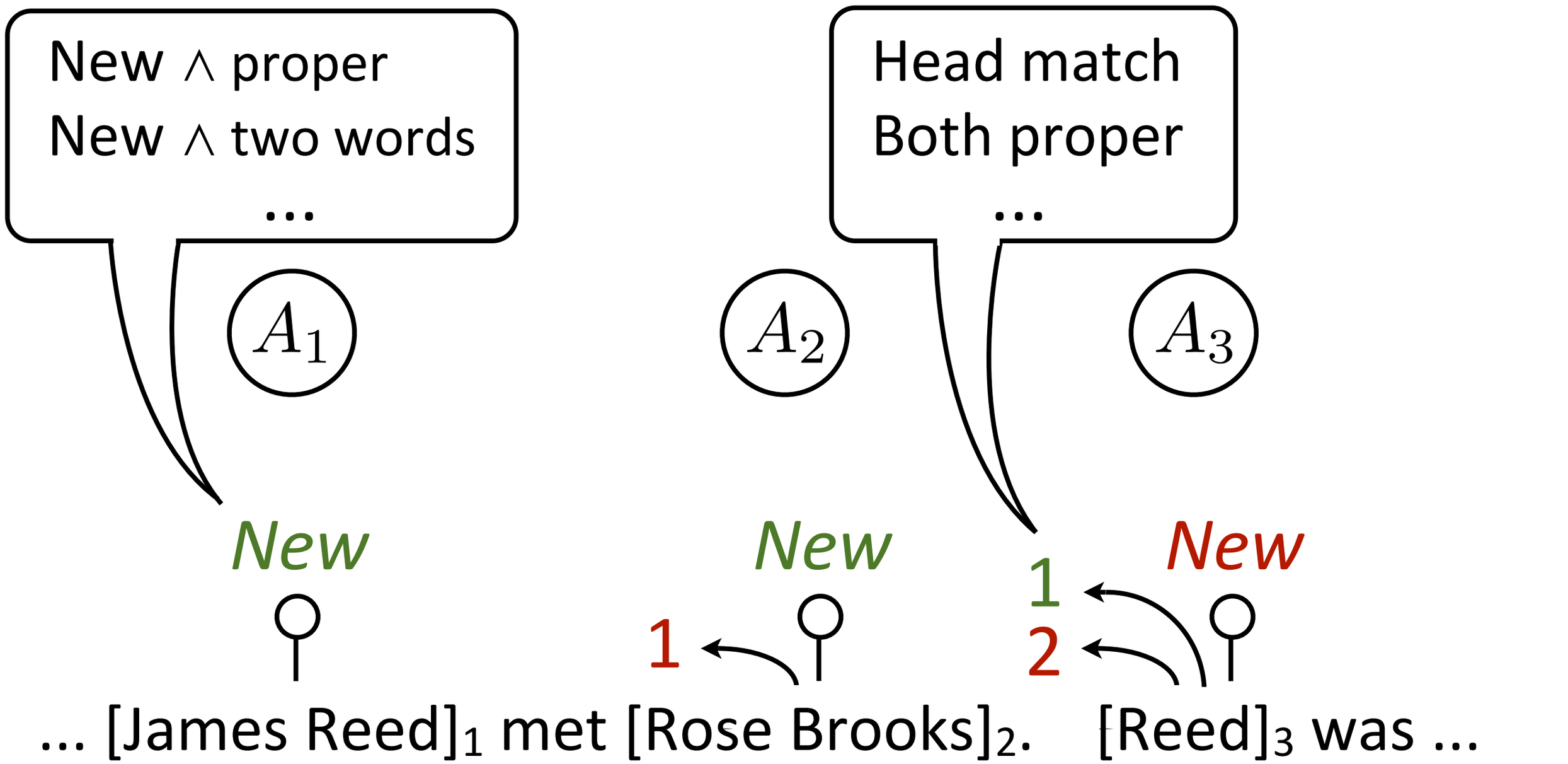
$$Pr(a_i|x) \propto \exp(w^T f(a_i, x))$$



[Denis and Baldridge (2008)]

Basic Model

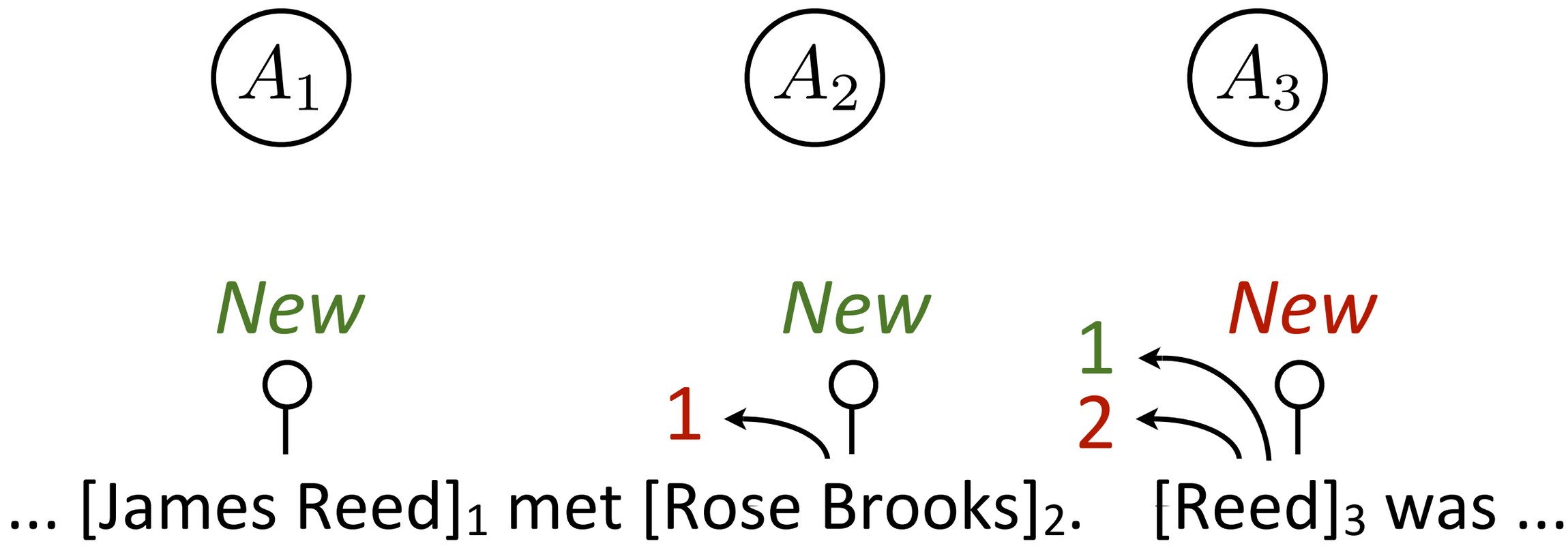
$$Pr(a_i|x) \propto \exp(w^T f(a_i, x))$$



[Denis and Baldridge (2008)]

Basic Model

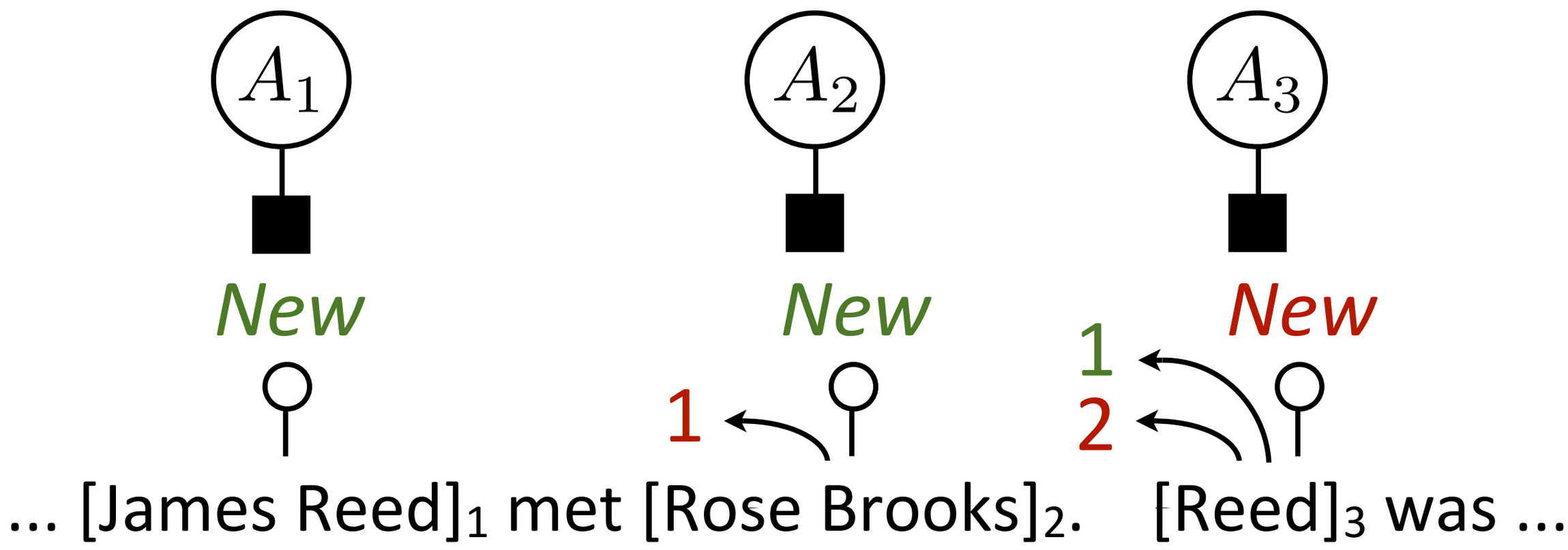
$$Pr(a_i|x) \propto \exp(w^T f(a_i, x))$$



[Denis and Baldridge (2008)]

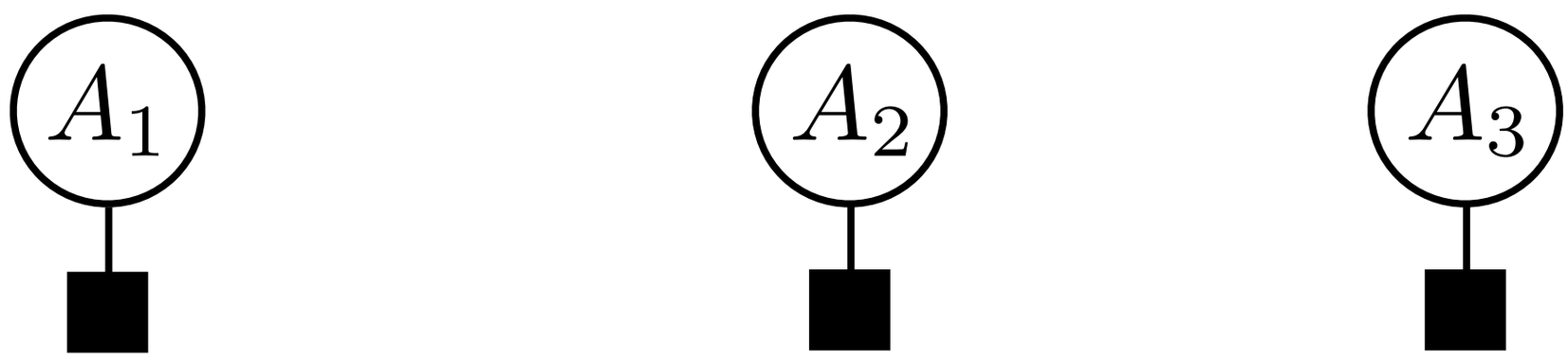
Basic Model

$$Pr(a_i|x) \propto \exp(w^T f(a_i, x))$$



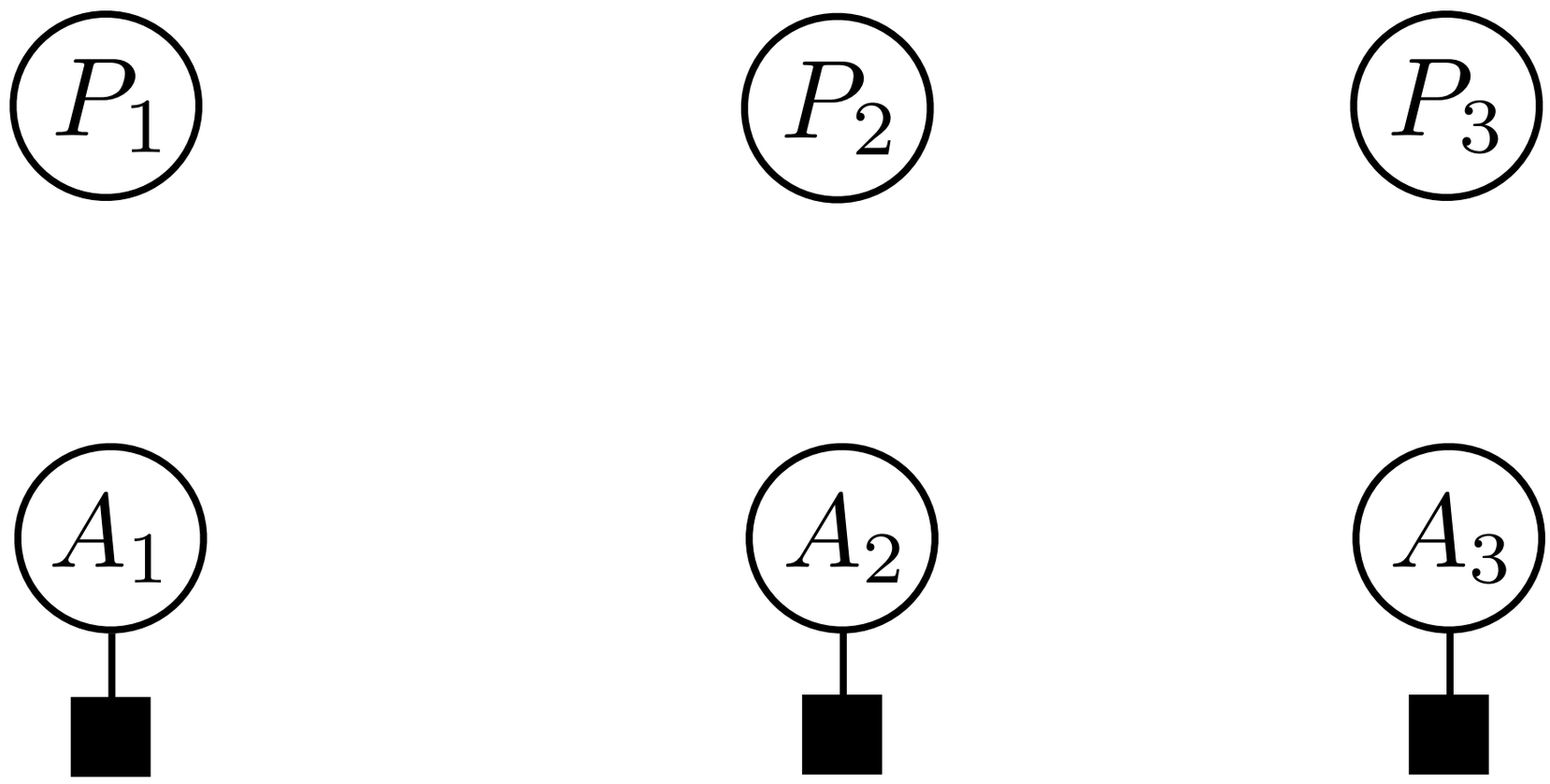
[Denis and Baldridge (2008)]

Decentralized Model



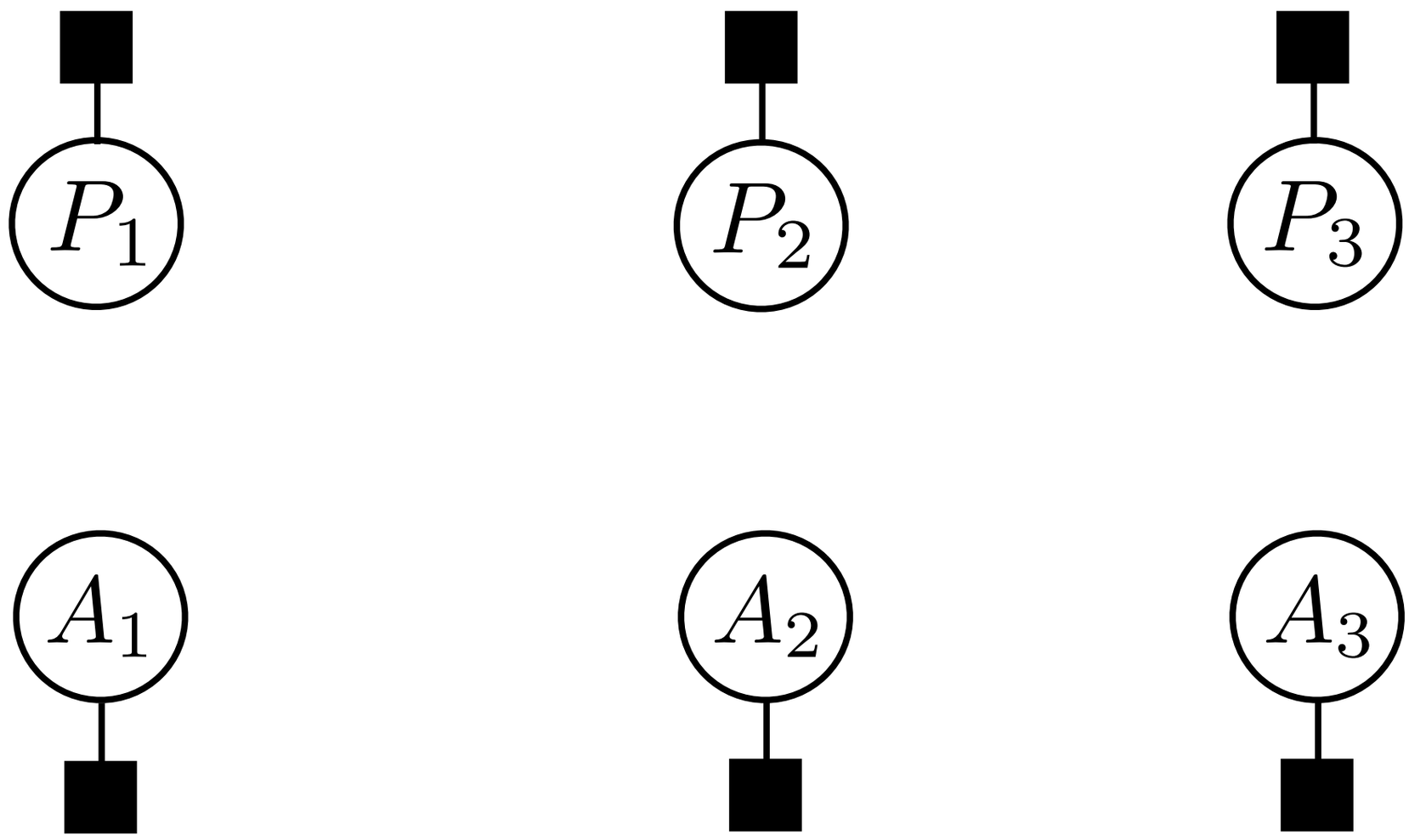
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



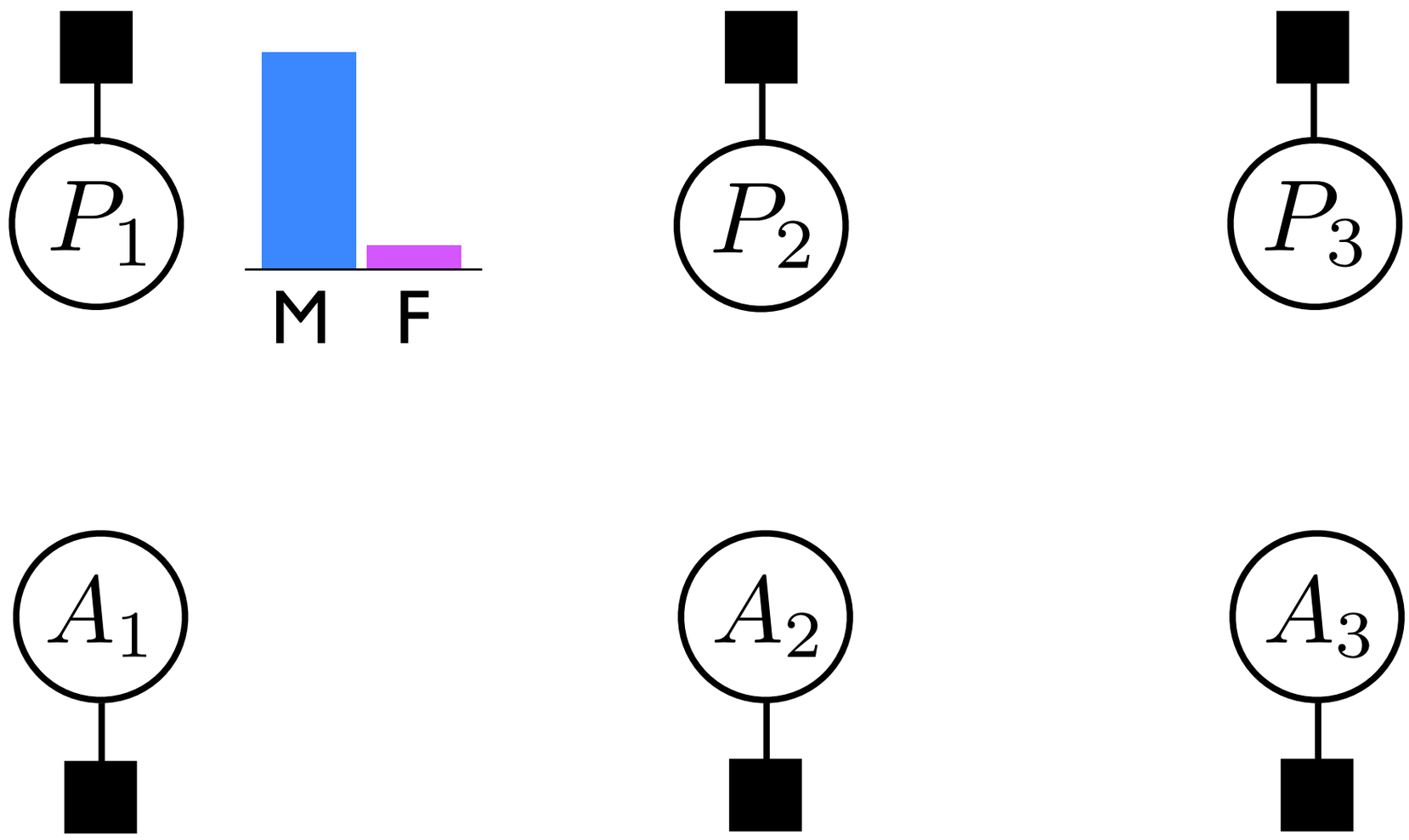
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

DECENTRALIZED Model



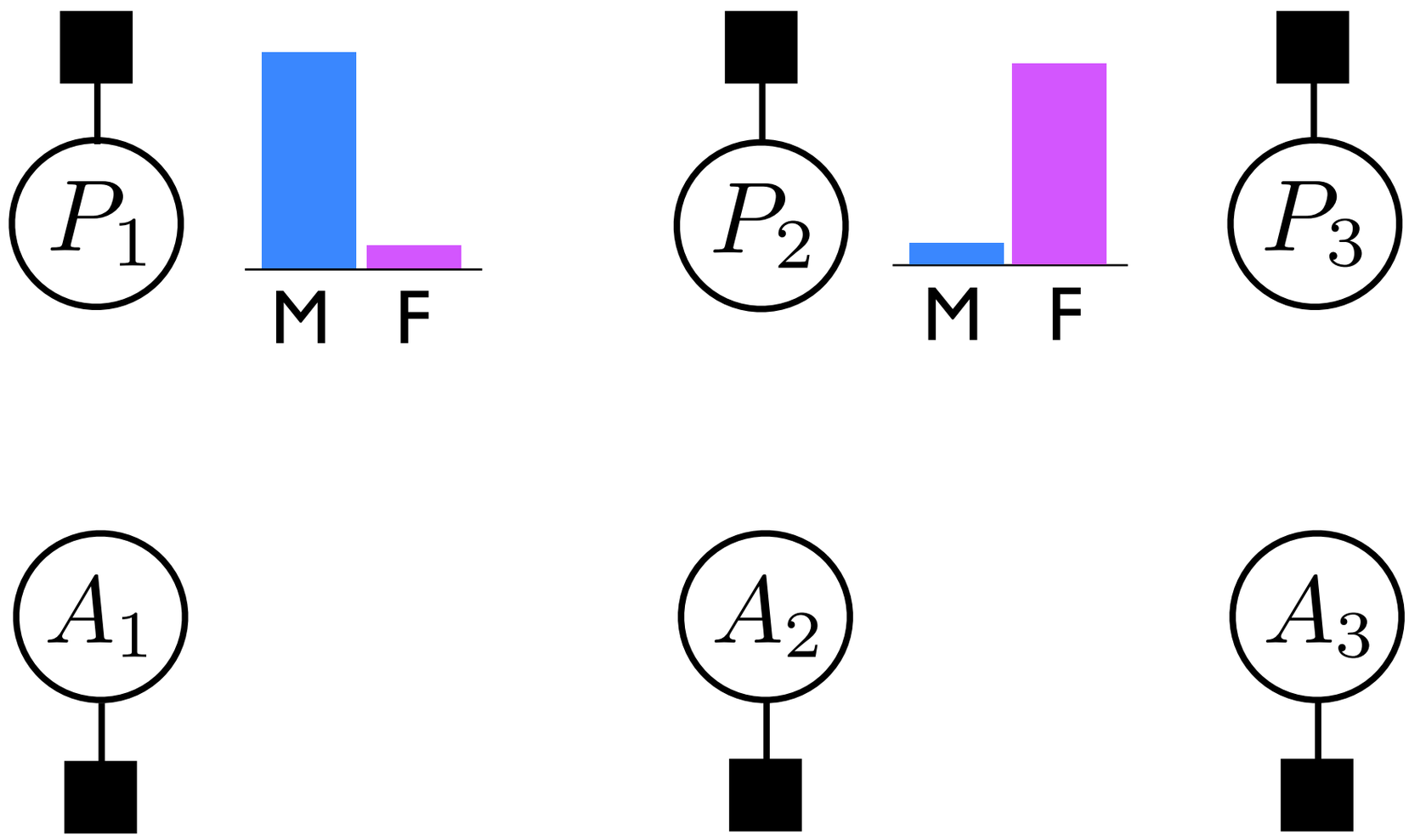
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



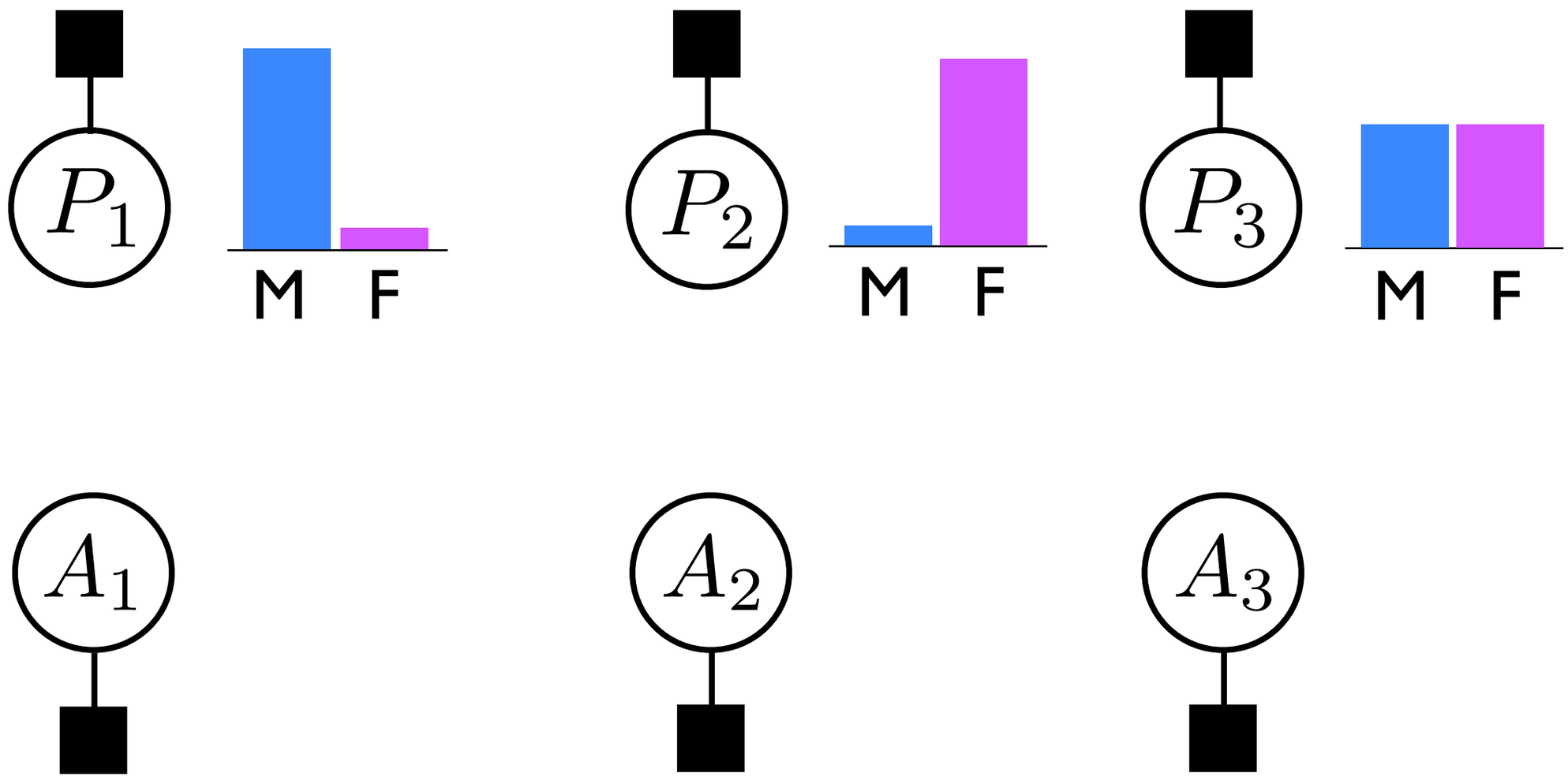
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



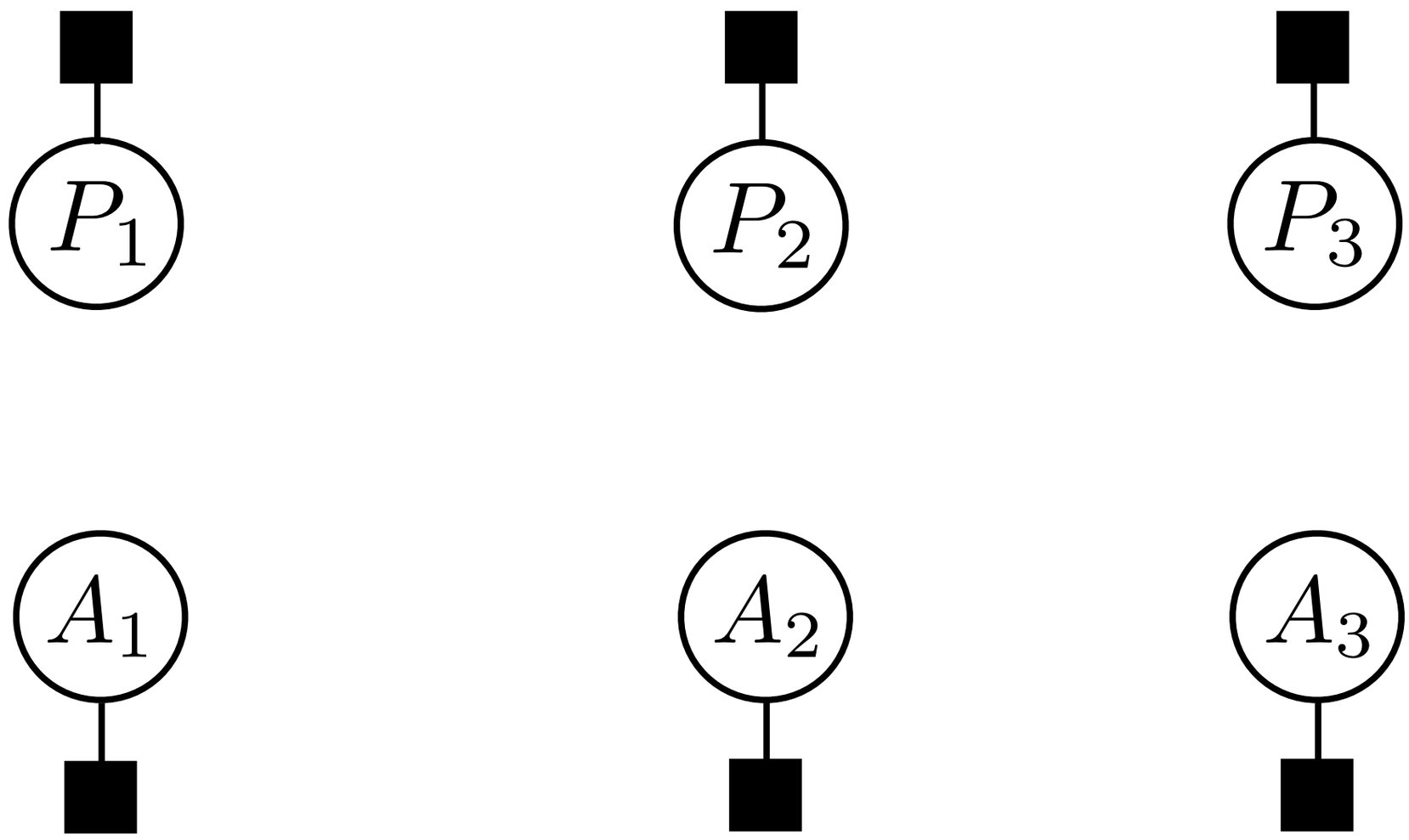
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



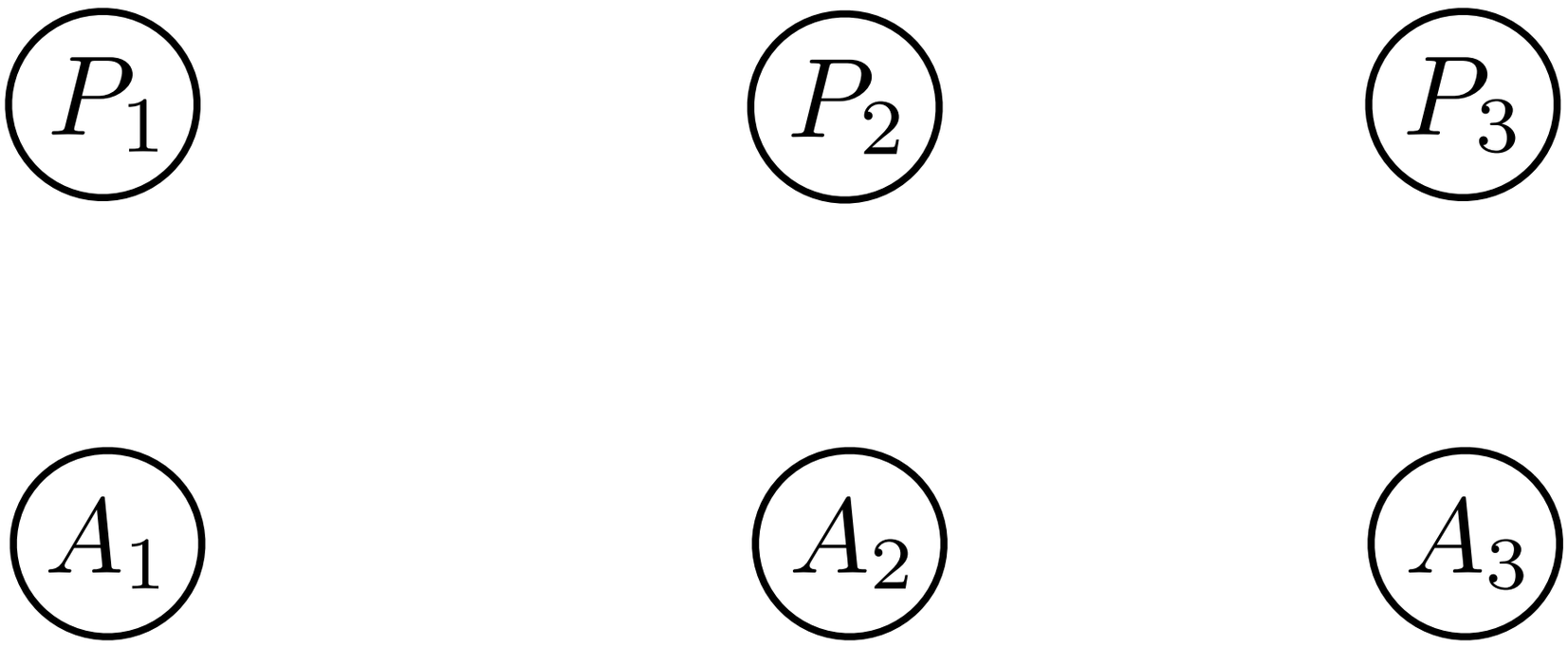
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

DECENTRALIZED Model



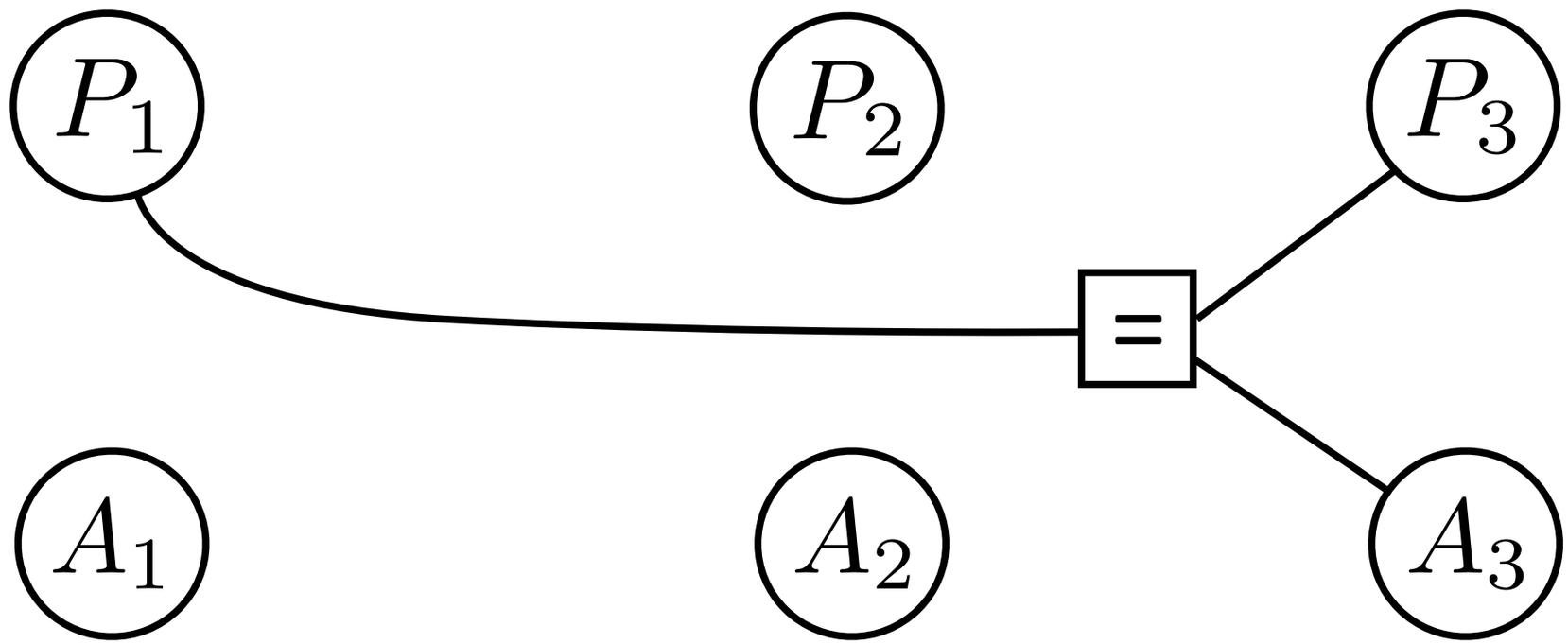
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



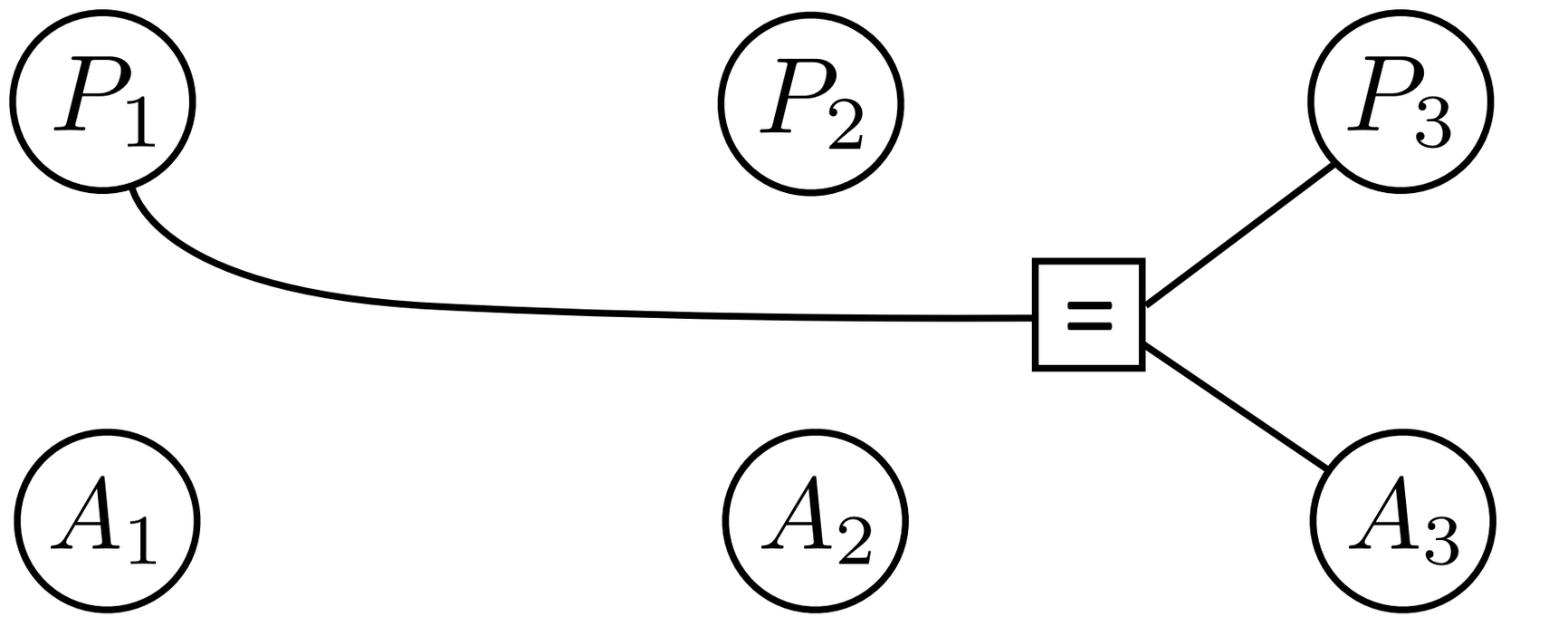
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

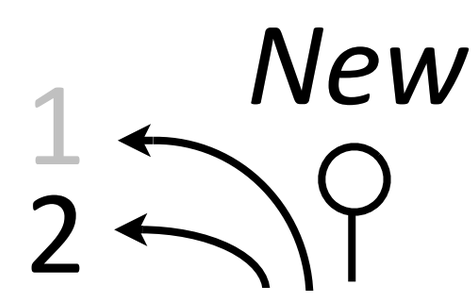
Decentralized Model



... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

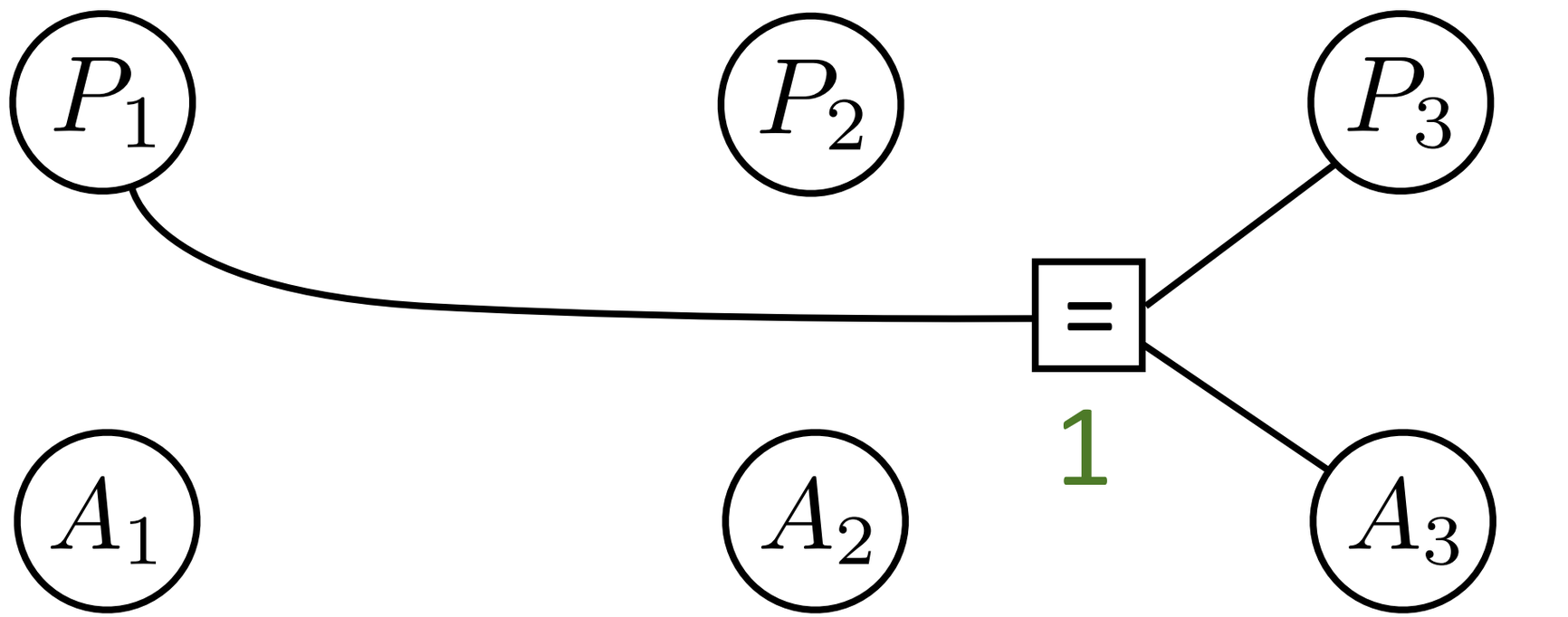
Decentralized Model

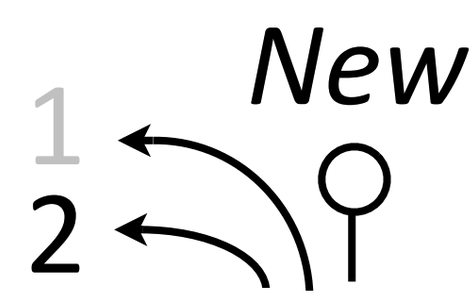




... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

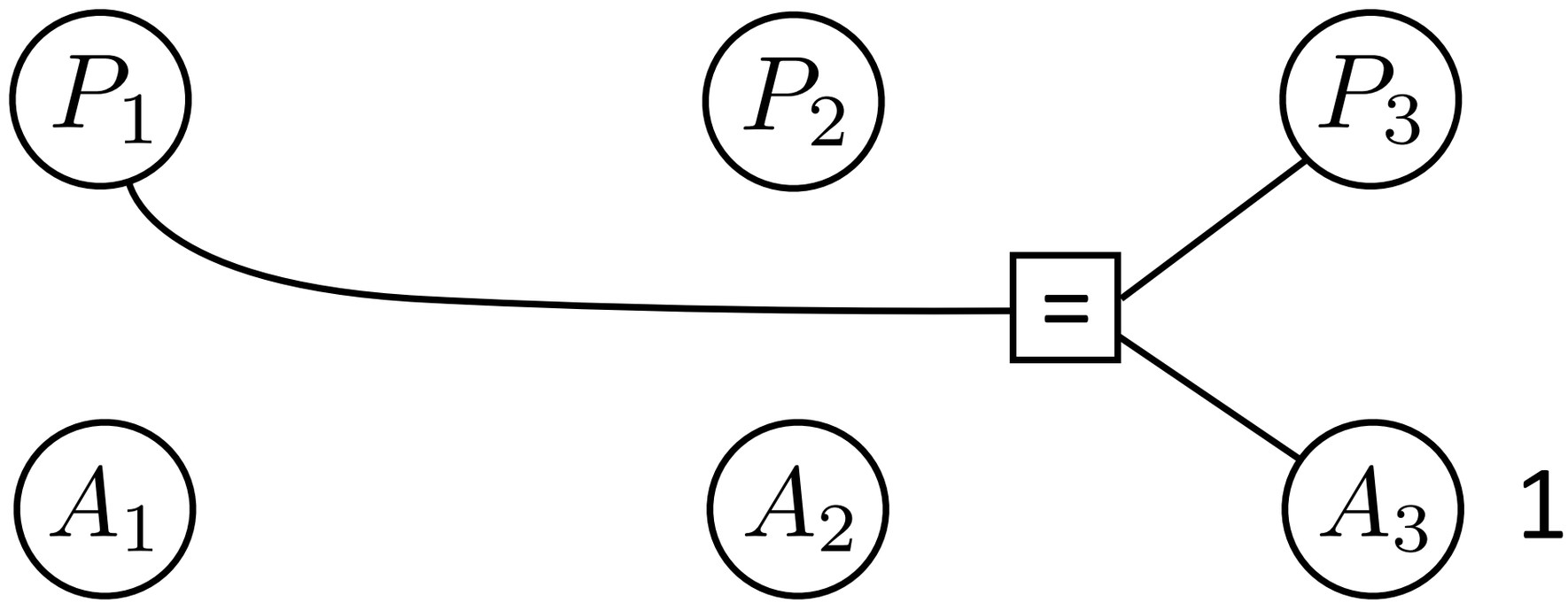
Decentralized Model

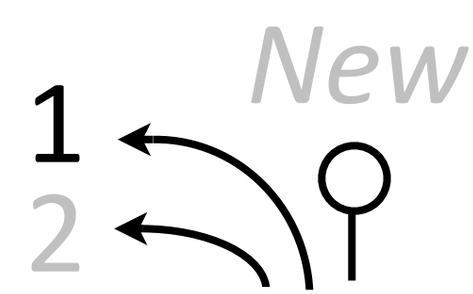




... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

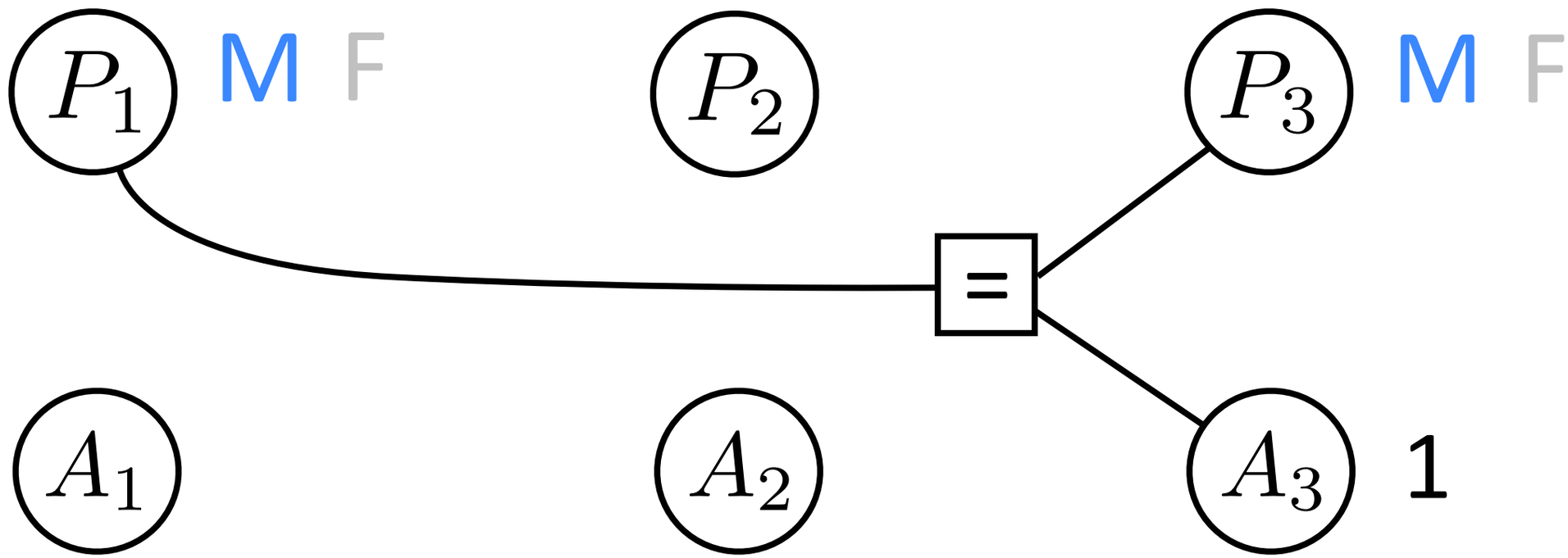
Decentralized Model

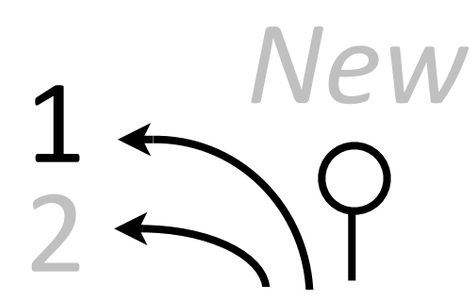




... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

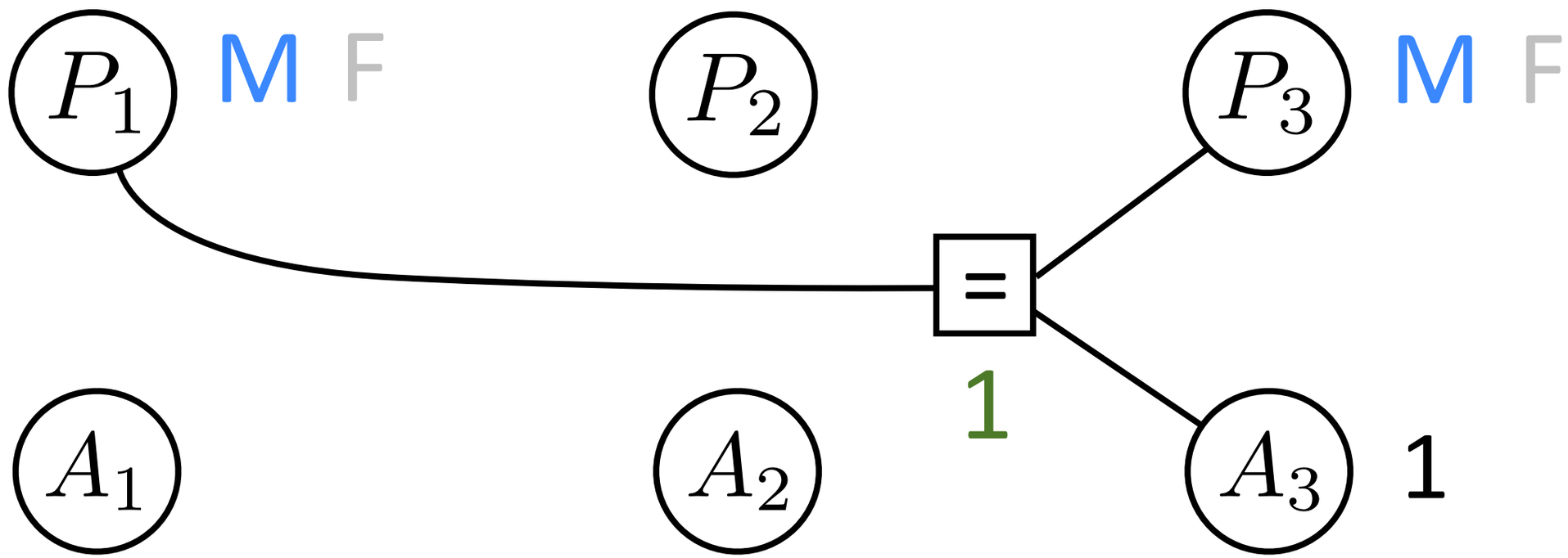
Decentralized Model

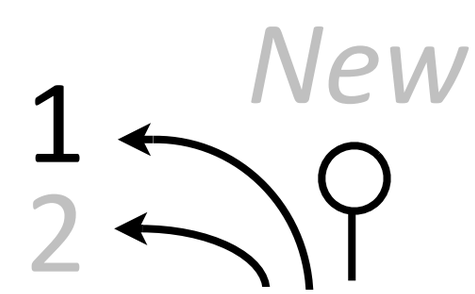




... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

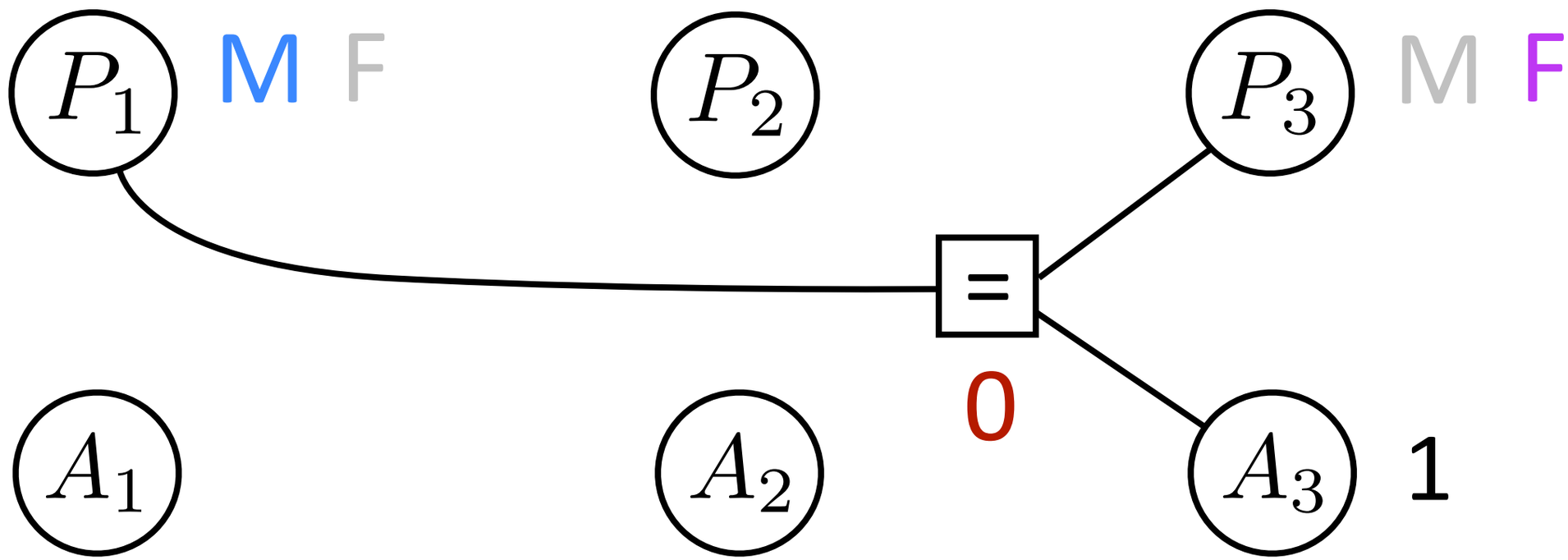
Decentralized Model

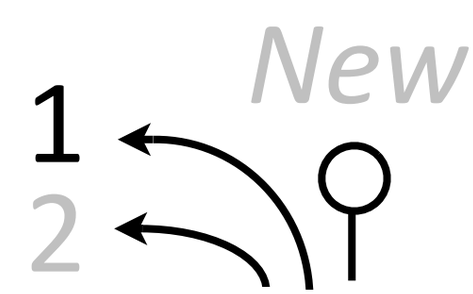




... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

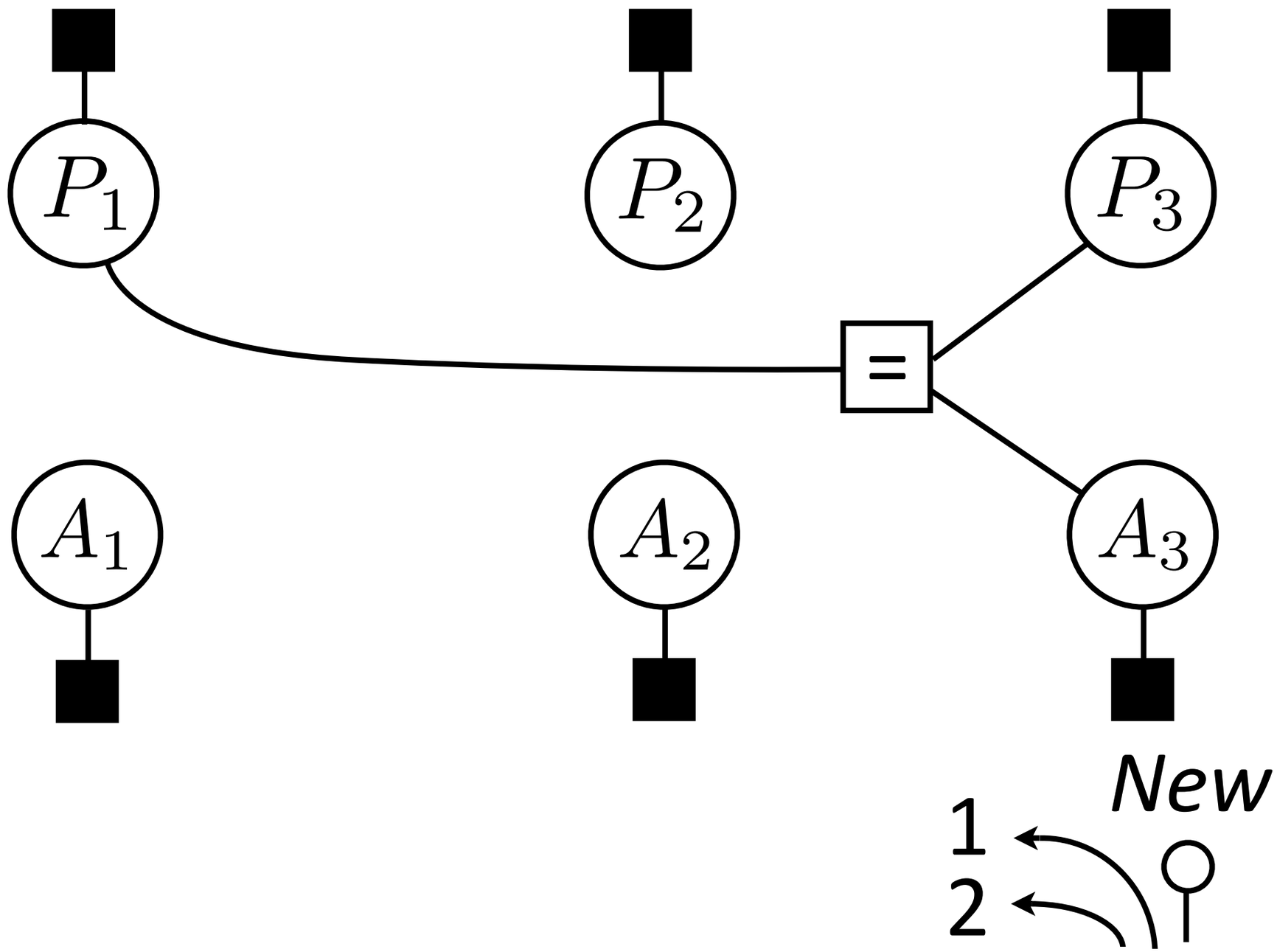
Decentralized Model





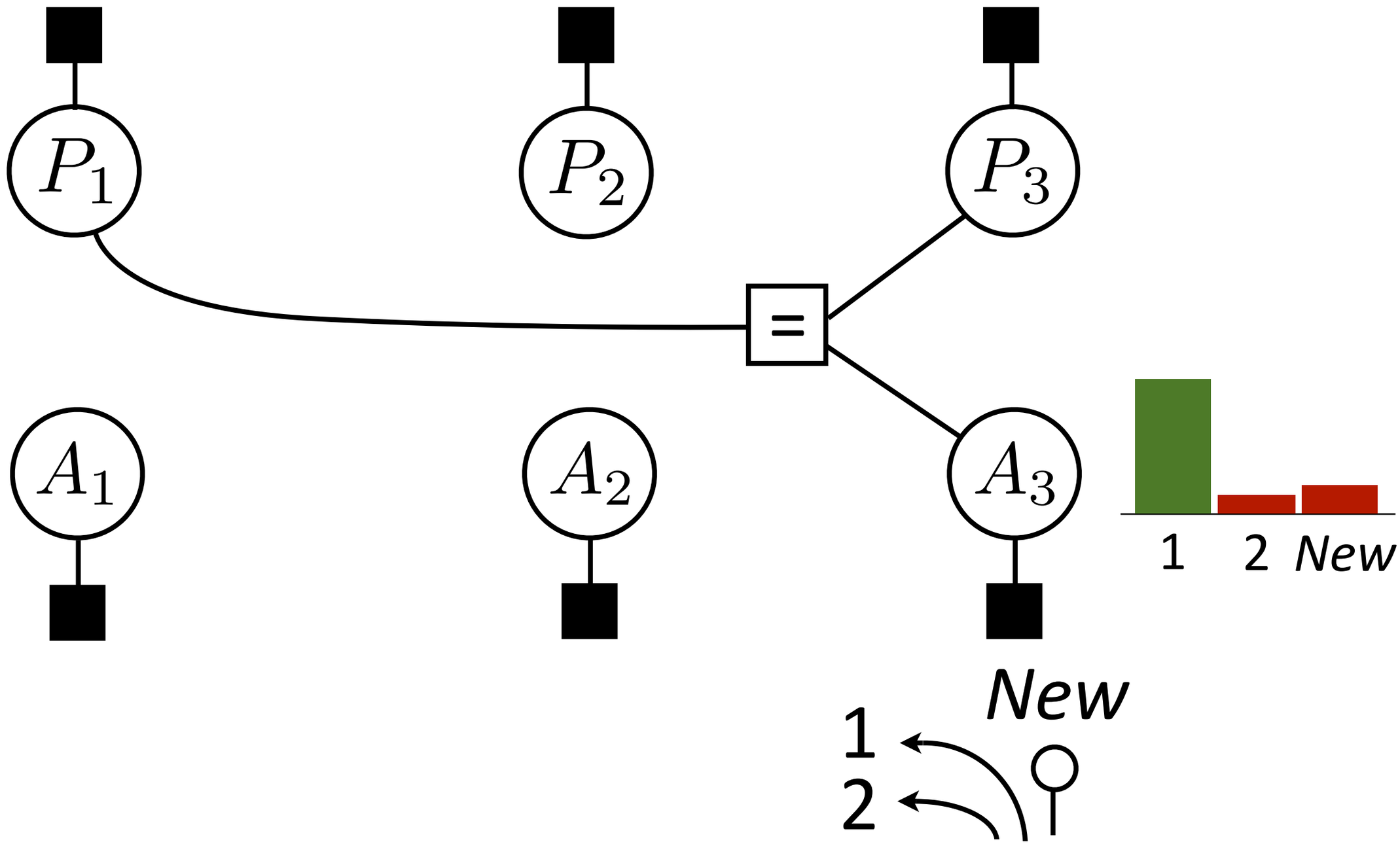
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



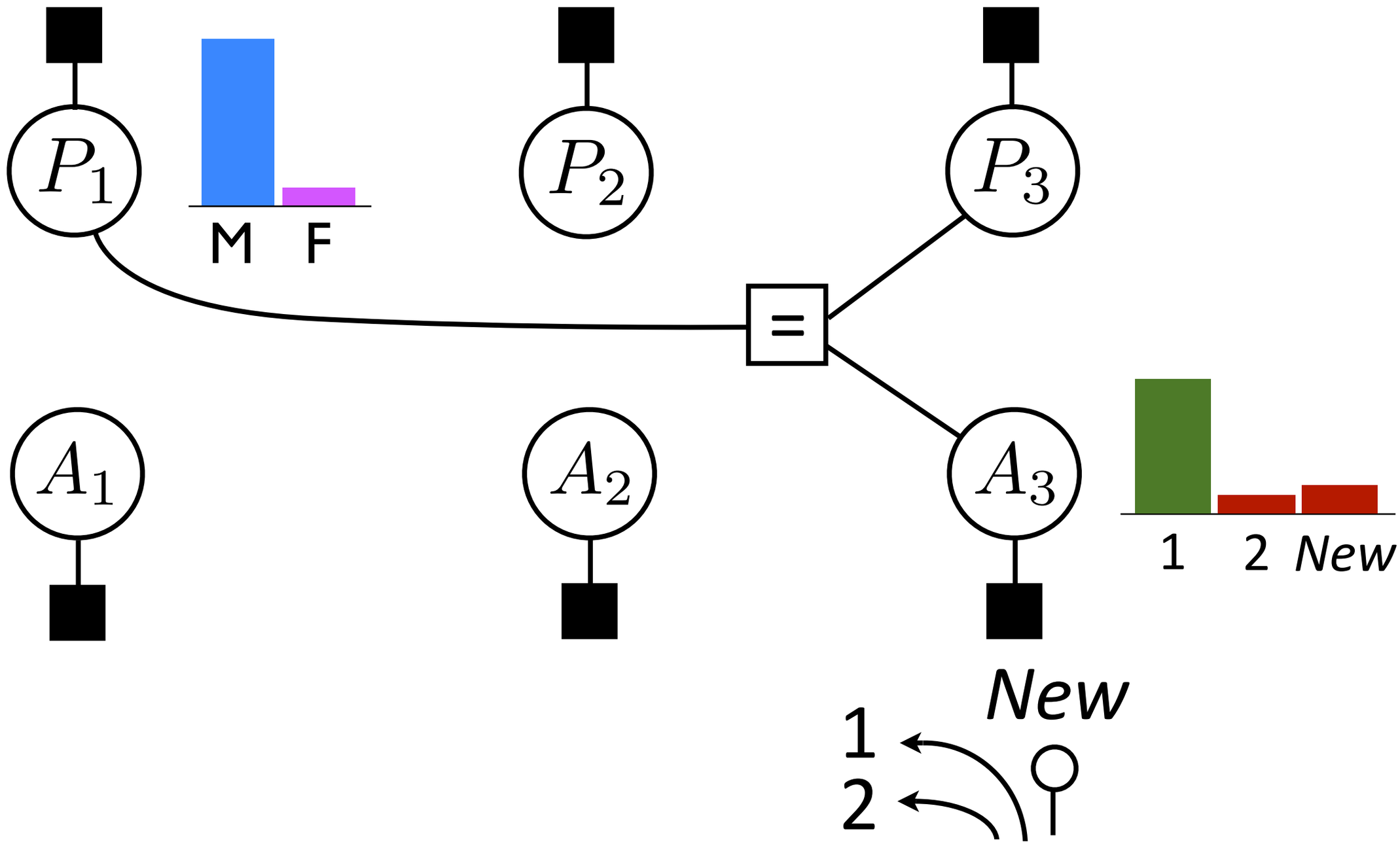
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

DECENTRALIZED Model



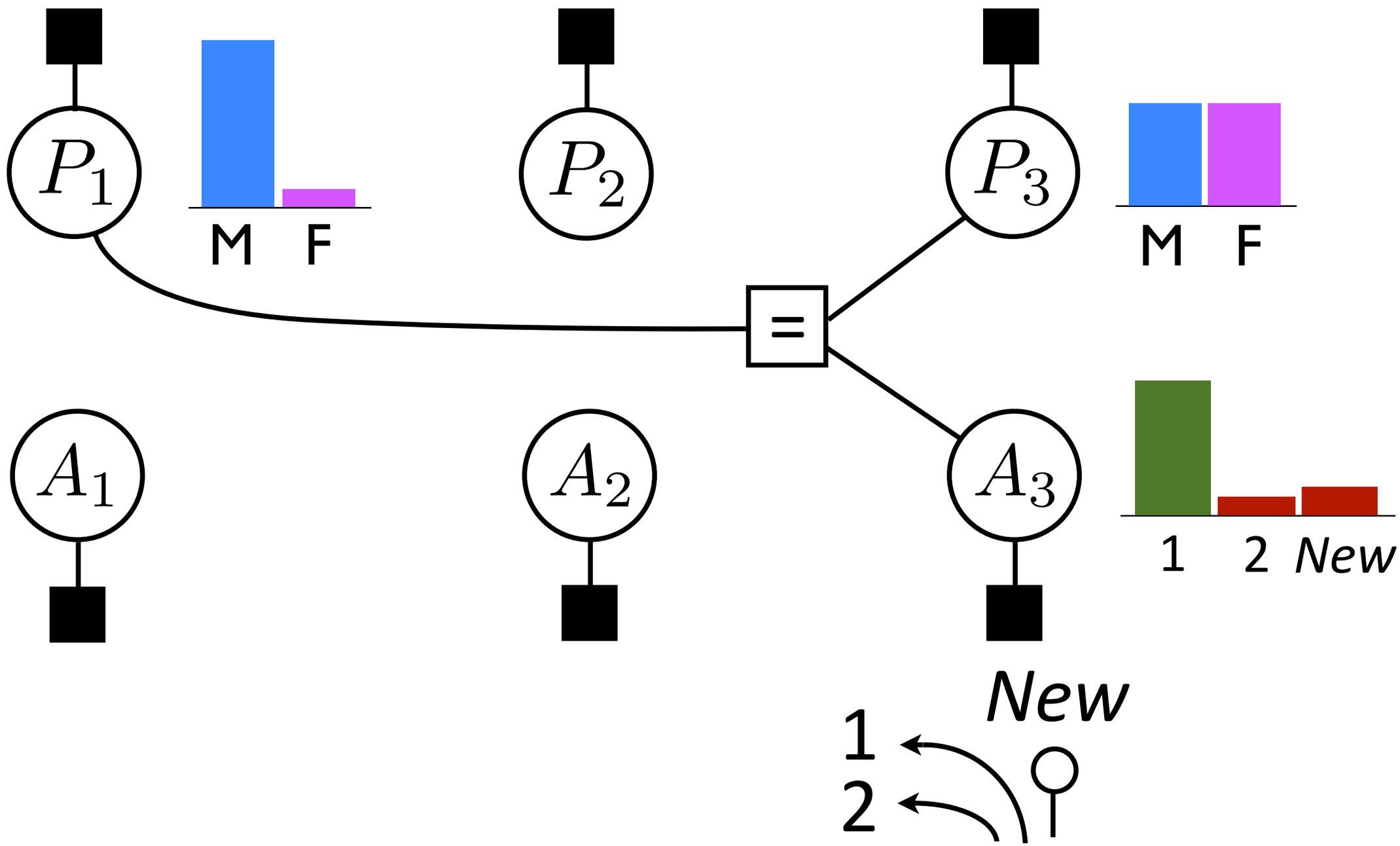
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



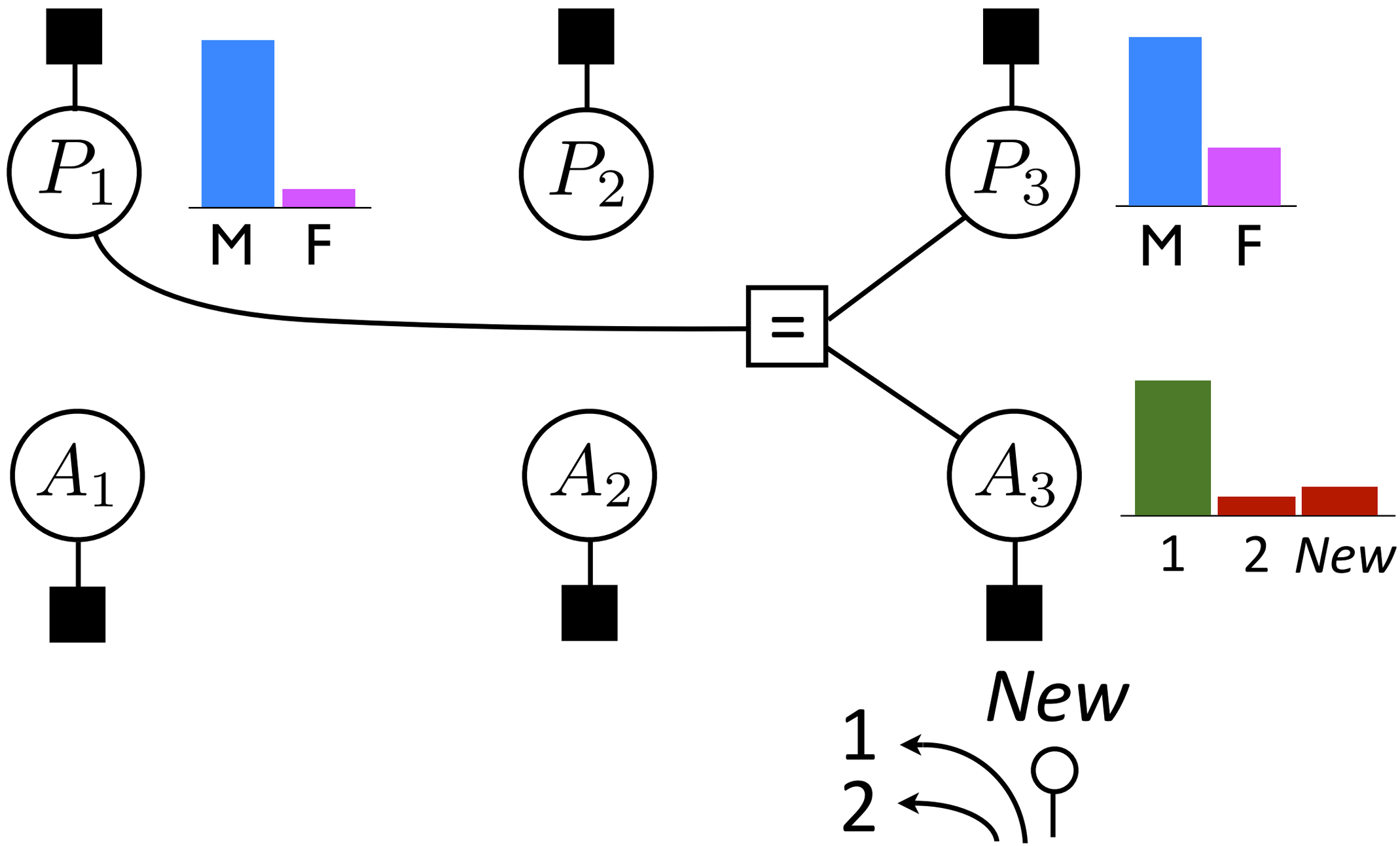
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model



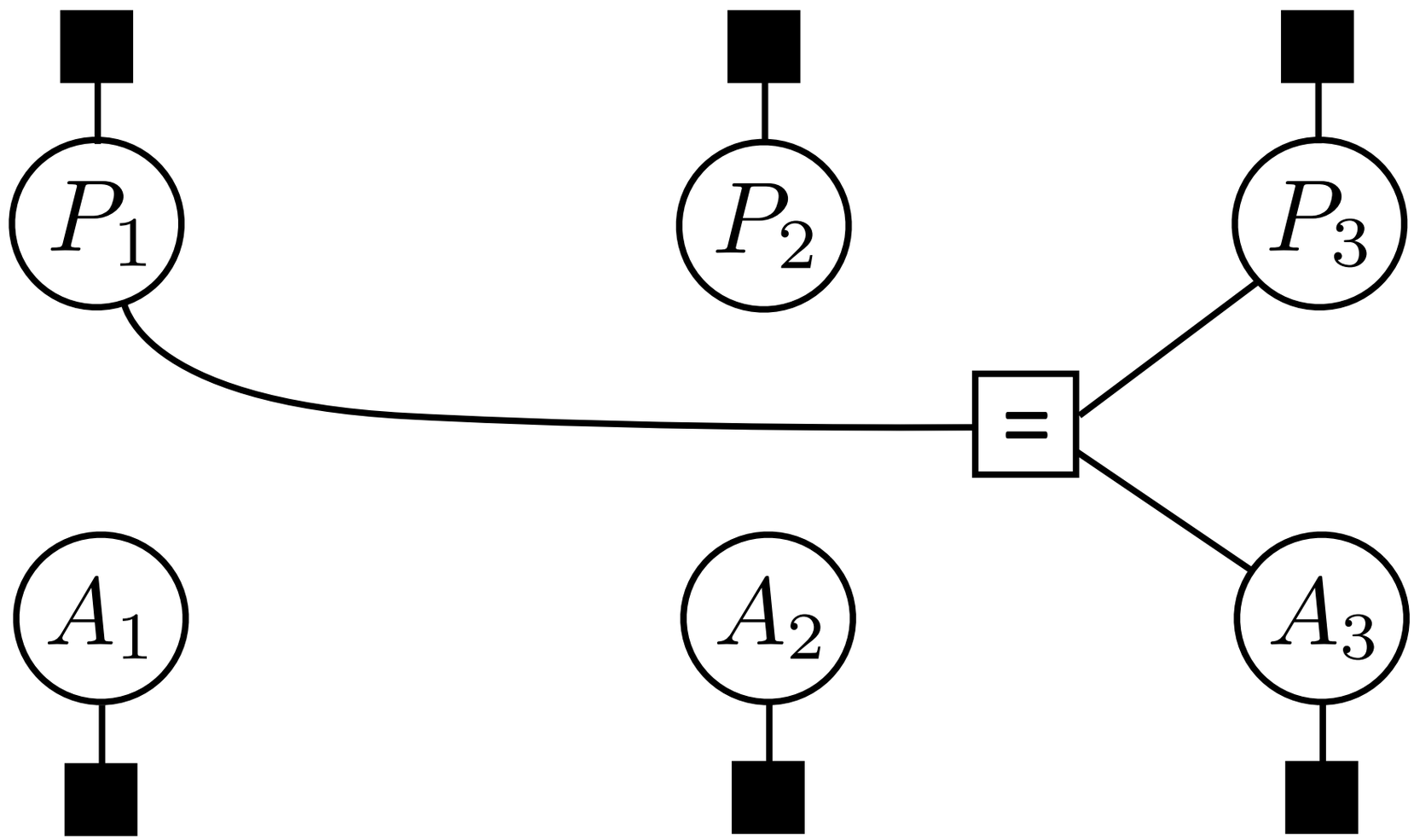
... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

Decentralized Model

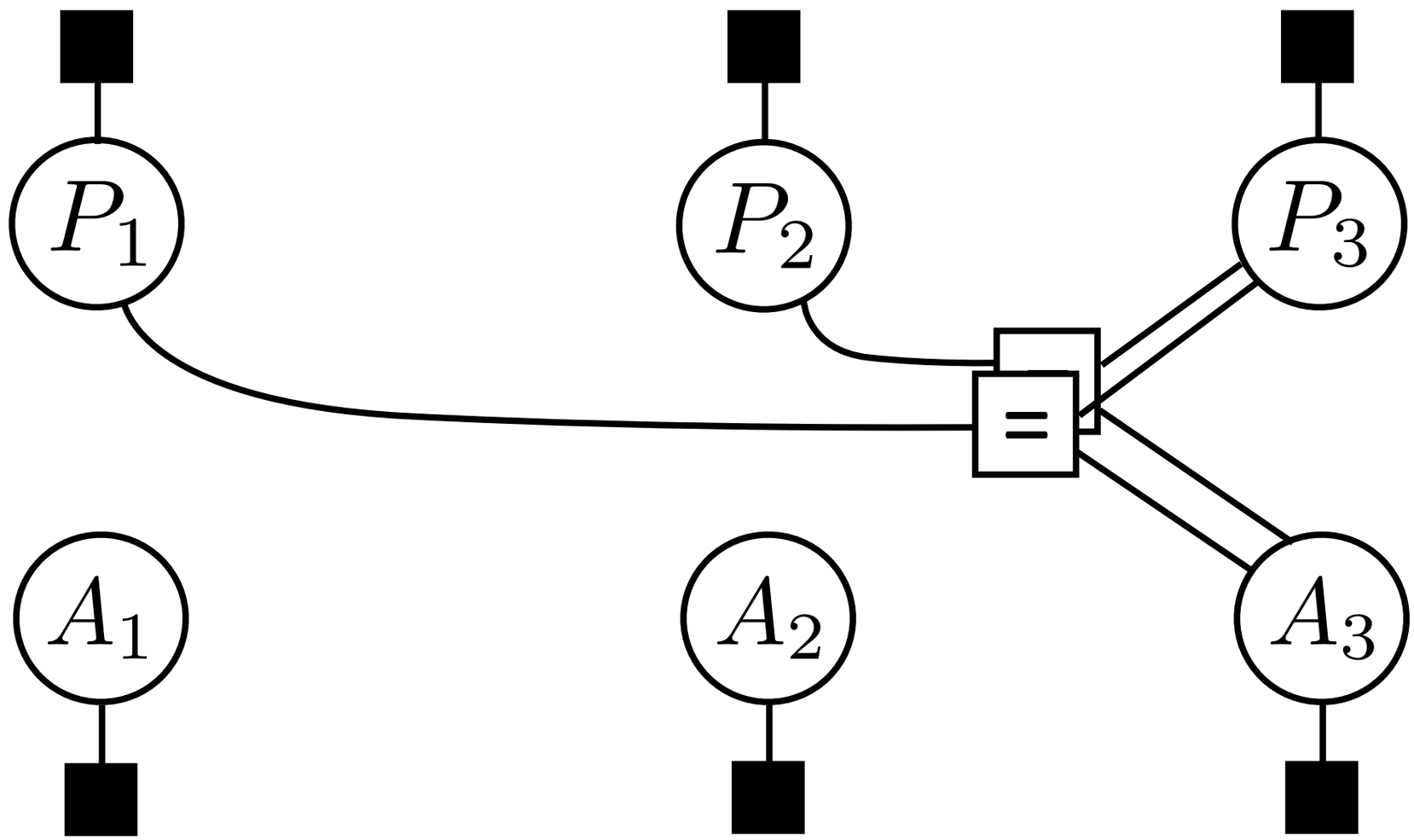


... [James Reed]$_1$ met [Rose Brooks]$_2$. [Reed]$_3$ was ...

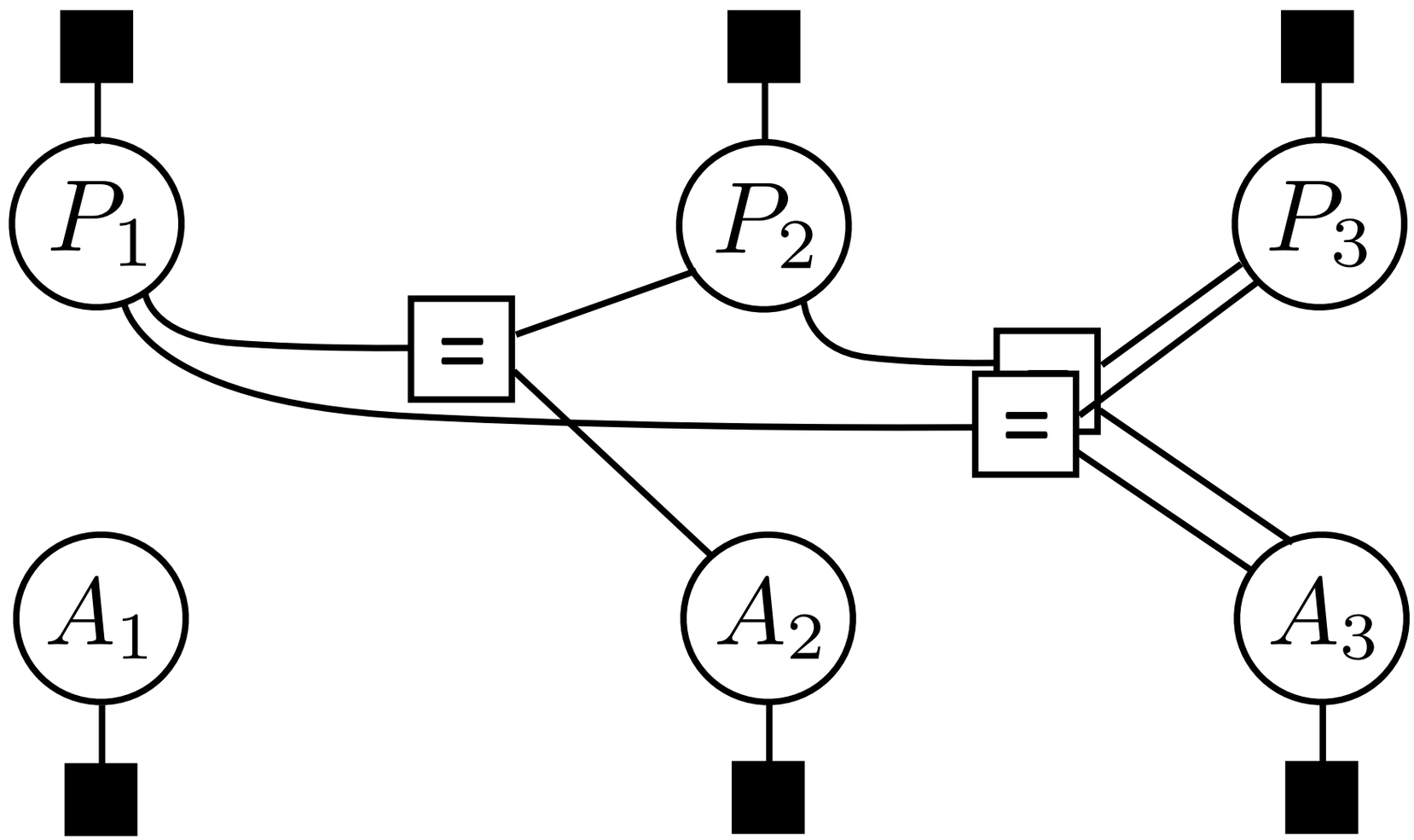
Decentralized Model



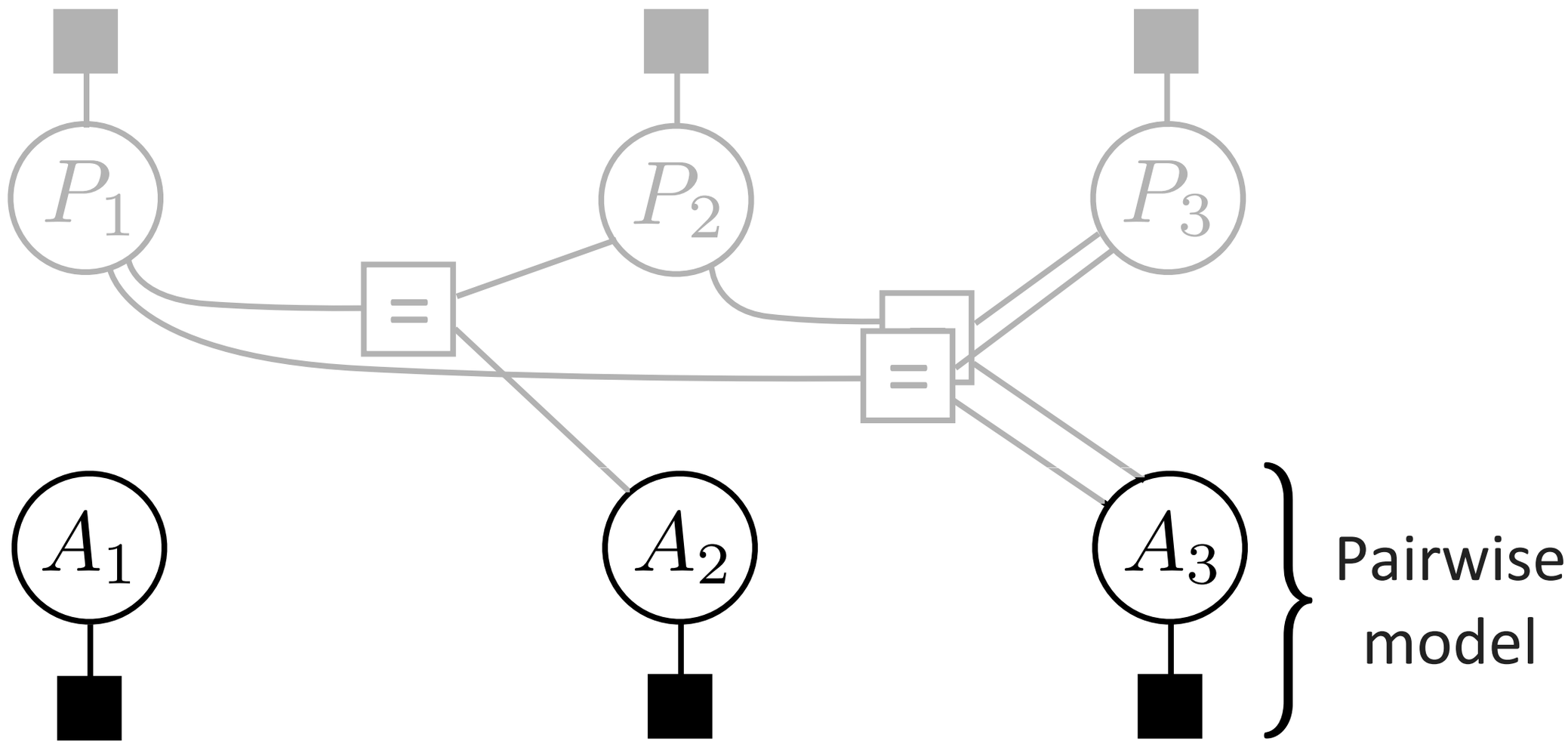
Decentralized Model



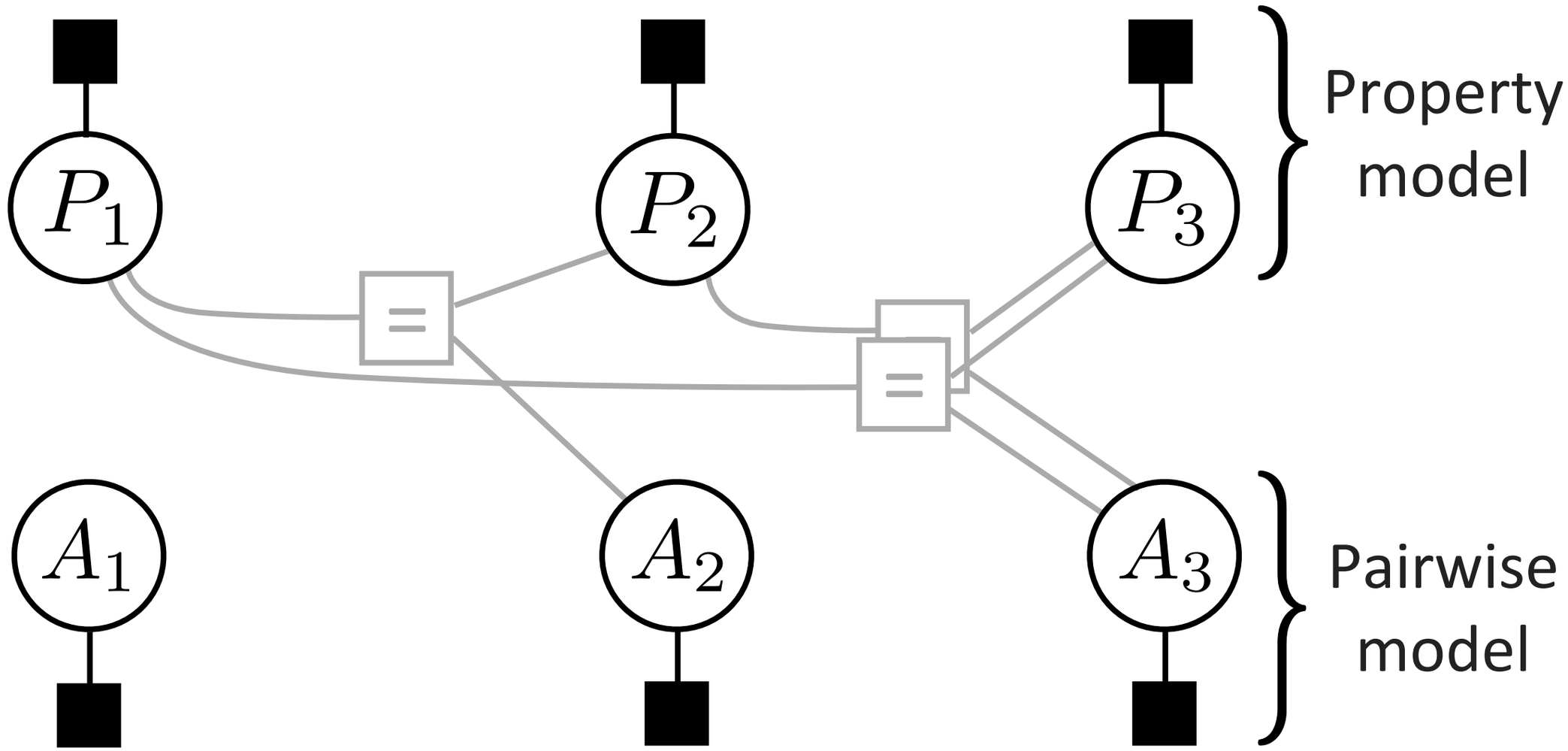
Decentralized Model



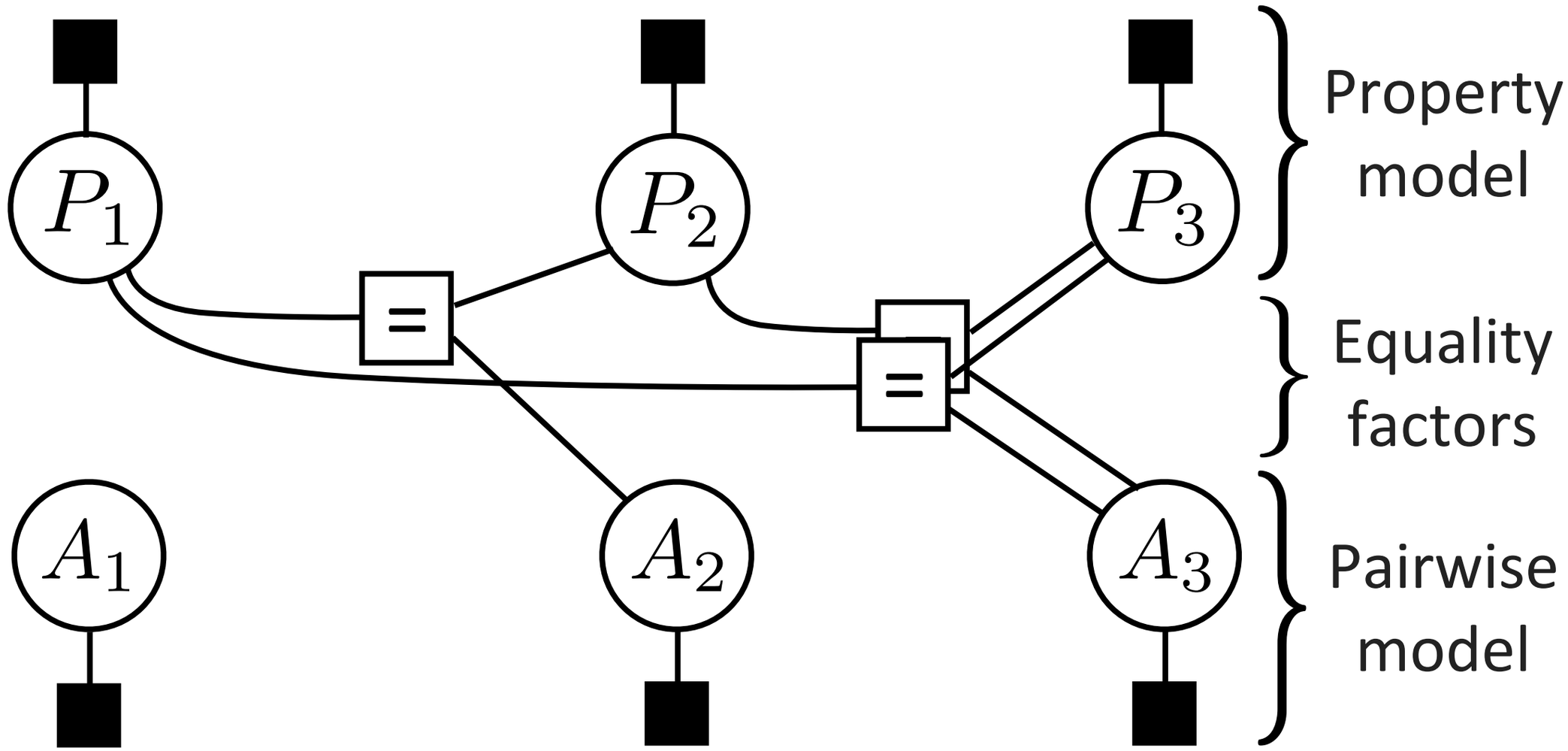
Decentralized Model



DECENTRALIZED Model



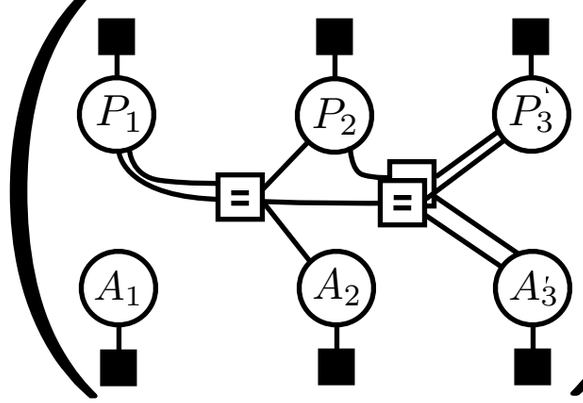
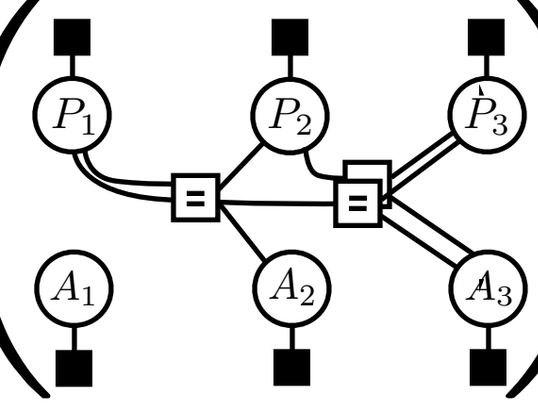
DECENTRALIZED Model



Inference

Inference

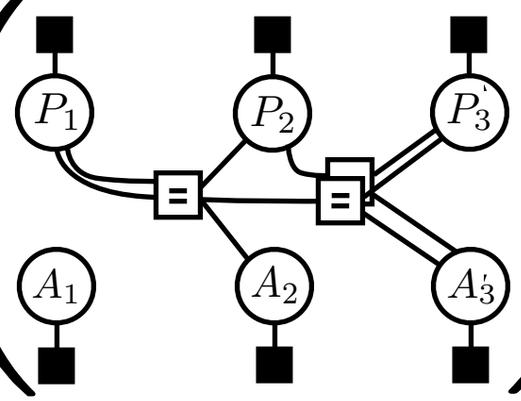
- Need to compute expected feature counts:

$\mathbb{E}_{\text{gold}} f\Bigg($  $\Bigg) - \mathbb{E}_{\text{all}} f\Bigg($  $\Bigg)$

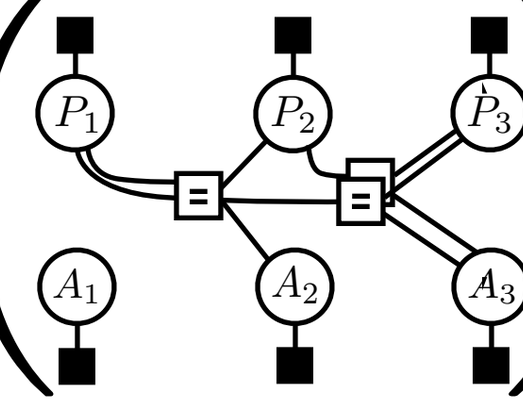
Inference

- Need to compute expected feature counts:

$$\mathbb{E}_{\text{gold}} f\Bigg($$



$$\Bigg) - \mathbb{E}_{\text{all}} f\Bigg($$



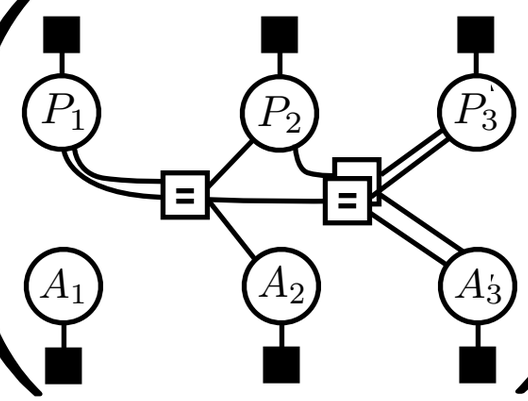
$$\Bigg)$$

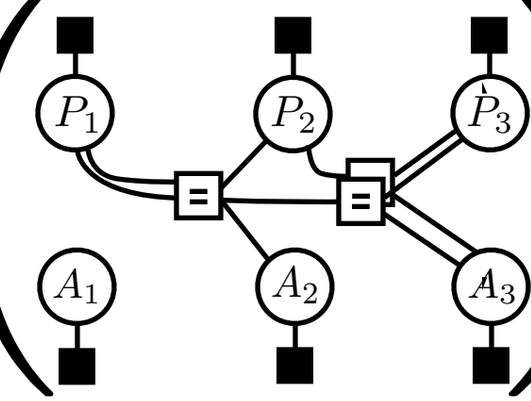
- Use belief propagation to compute marginals over variables

Inference

- Need to compute expected feature counts:

$$\mathbb{E}_{\text{gold}} f\Bigg(\Bigg. \Bigg. - \mathbb{E}_{\text{all}} f\Bigg(\Bigg. \Bigg.$$





- Use belief propagation to compute marginals over variables

- Decoding: max over each A_i marginal

Learning

Learning

- Optimize conditional log likelihood of training data

Learning

- Optimize conditional log likelihood of training data

$$\sum_i \log\left(Pr(a_g^i | x^i)\right)$$

Learning

- Optimize conditional log likelihood of training data

$$\sum_i \log\left(Pr(a_g^i | x^i)\right)$$

Training examples

Learning

- Optimize conditional log likelihood of training data

Gold antecedent vector

$$\sum_{i} \log \left(Pr(a_g^i | x^i) \right)$$

Training examples

Learning

- Optimize conditional log likelihood of training data

Gold antecedent vector

$$\sum_{i} \log \left(Pr(a_g^i | x^i) \right)$$

Training examples

Observed document properties

Learning

- Optimize conditional log likelihood of training data

$$\sum_i \log \left(\sum_{a_g^i \in \mathcal{A}(C)} Pr(a_g^i | x^i) \right)$$

Learning

- Optimize conditional log likelihood of training data

$$\sum_i \log \left(\sum_{a_g^i \in \mathcal{A}(C)} Pr(a_g^i | x^i) \right)$$

Antecedent choices consistent with gold standard

Learning

Learning

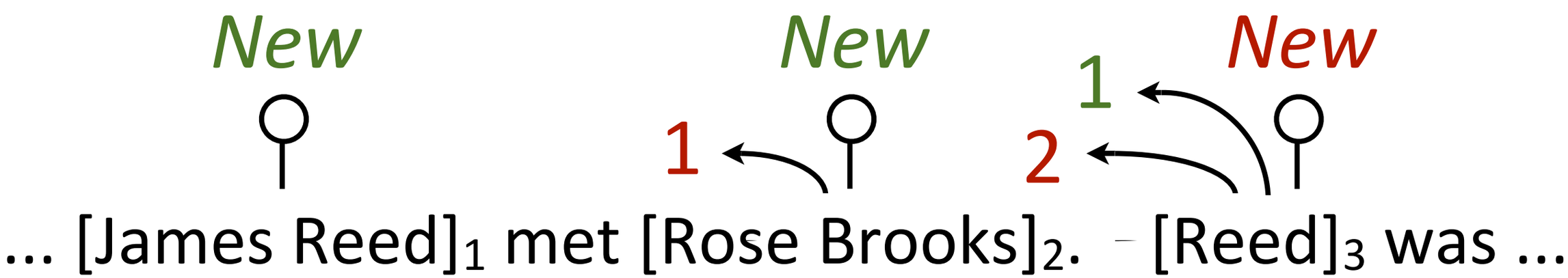
- Want to optimize for MUC, B^3, CEAF, etc.

Learning

- Want to optimize for MUC, B^3, CEAF, etc.
- Use a decomposable metric as a proxy

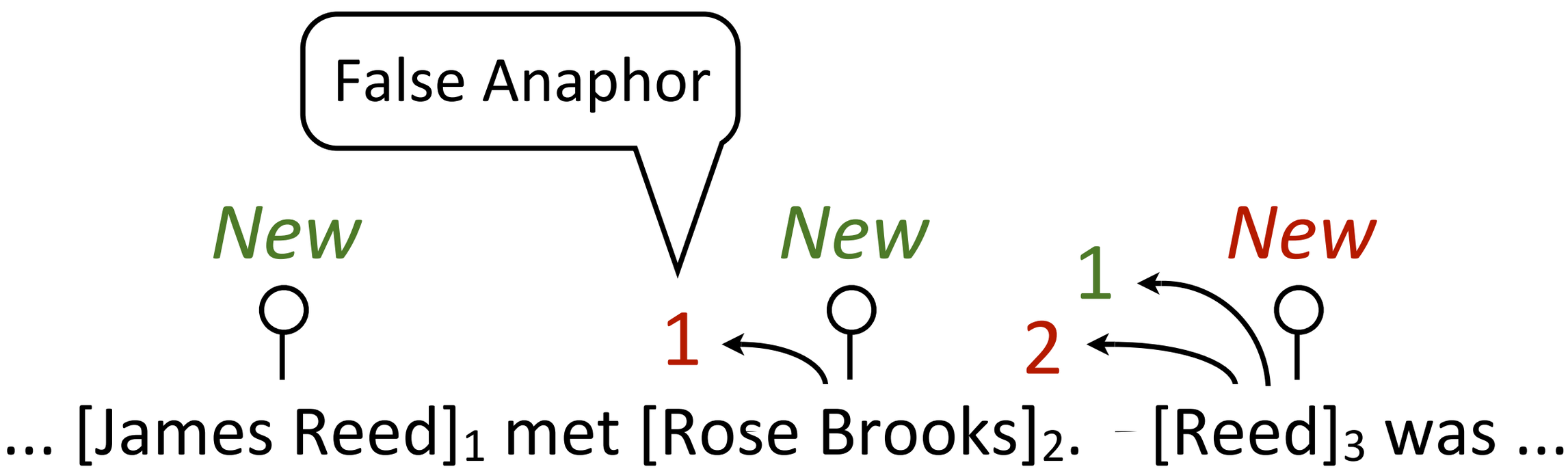
Learning

- Want to optimize for MUC, B^3, CEAF, etc.
- Use a decomposable metric as a proxy



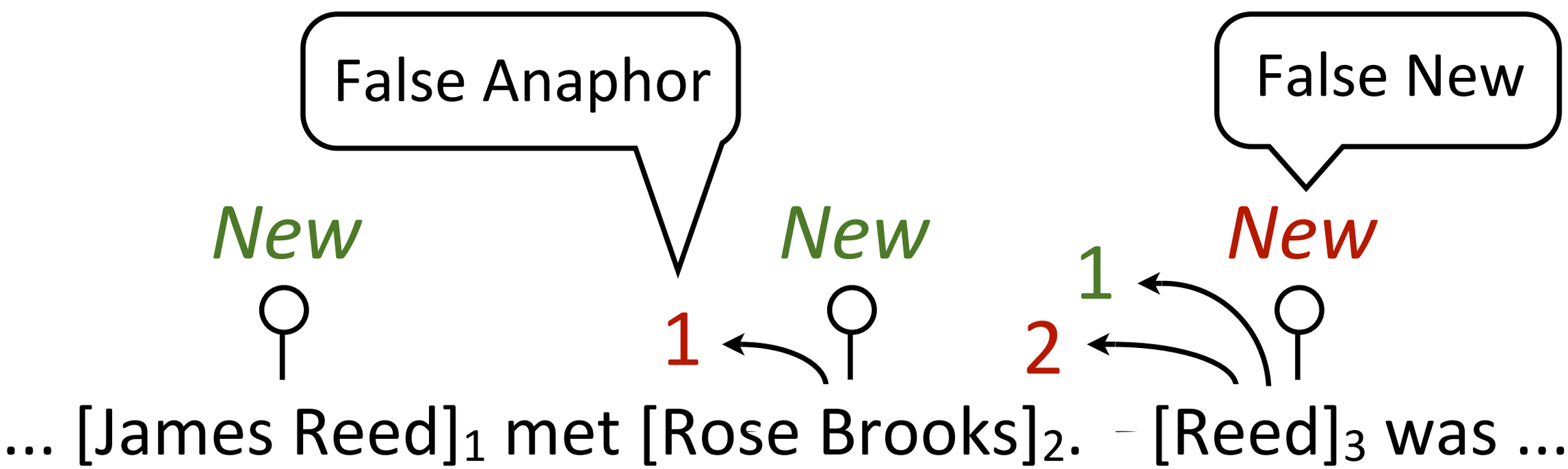
Learning

- Want to optimize for MUC, B^3, CEAF, etc.
- Use a decomposable metric as a proxy



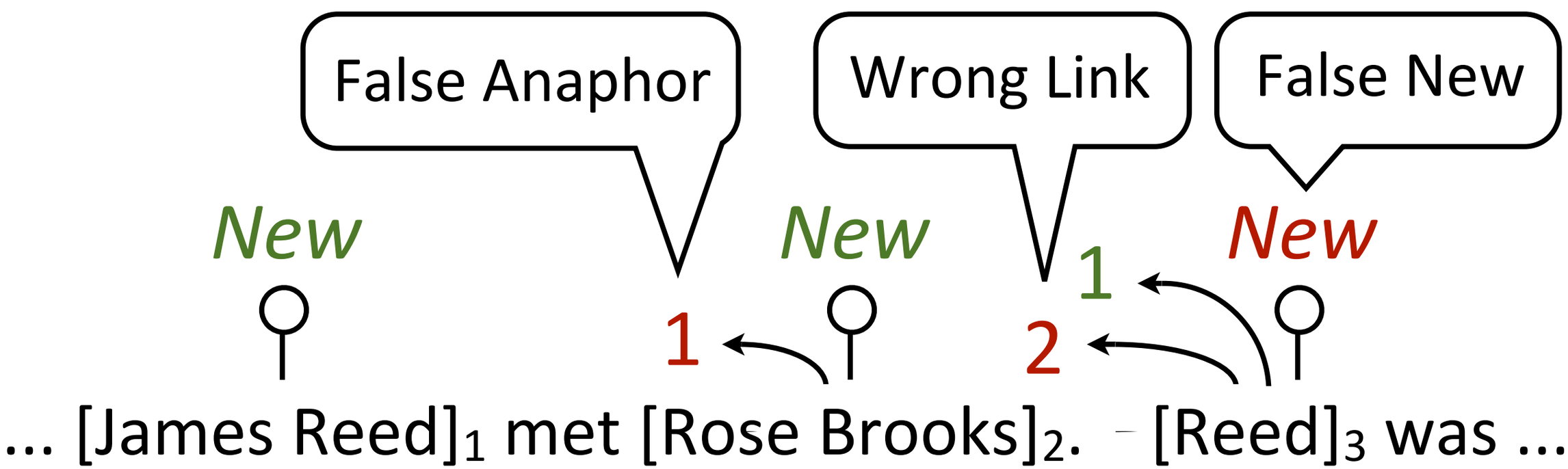
Learning

- Want to optimize for MUC, B^3, CEAF, etc.
- Use a decomposable metric as a proxy



Learning

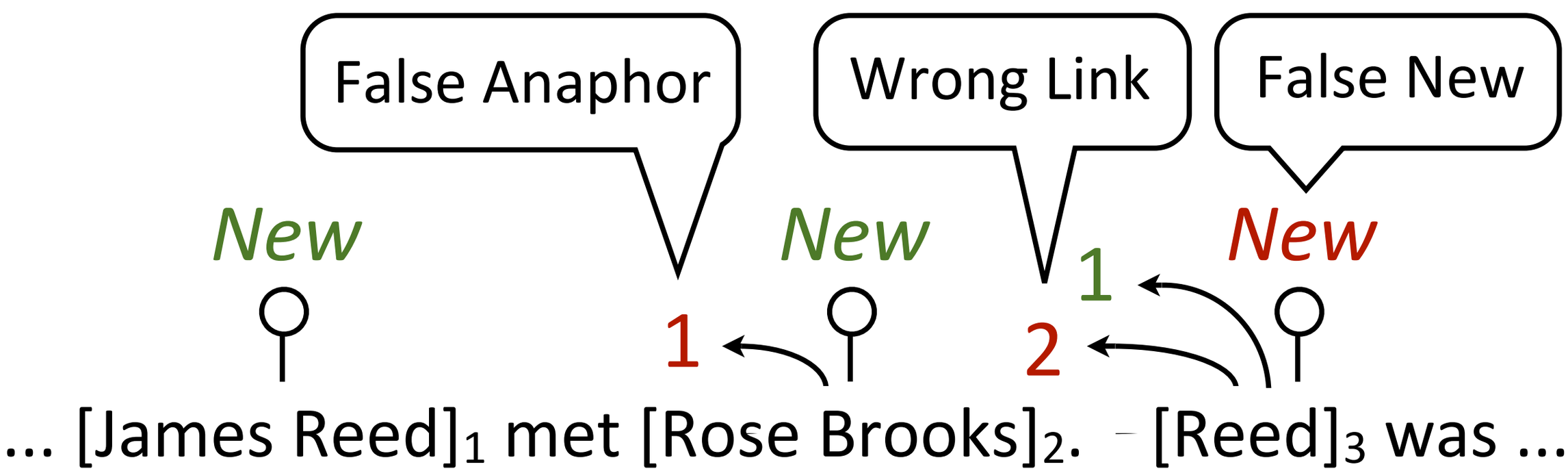
- Want to optimize for MUC, B^3, CEAF, etc.
- Use a decomposable metric as a proxy



Learning

- Want to optimize for MUC, B^3, CEAF, etc.
- Use a decomposable metric as a proxy

$$k_1(\text{False Anaphors}) + k_2(\text{False News}) + k_3(\text{Wrong Links})$$



Learning

$$\sum_i \log \left(\sum_{a_g^i \in \mathcal{A}(C)} Pr(a_g^i | x^i) \right)$$

Learning

- Incorporate this loss with *softmax-margin* by adding it as a feature to the pairwise model

$$\sum_i \log \left(\sum_{a_g^i \in \mathcal{A}(C)} Pr(a_g^i | x^i) \right)$$

[Gimpel and Smith (2010)]

Learning

- Incorporate this loss with *softmax-margin* by adding it as a feature to the pairwise model

$$\sum_i \log \left(\sum_{a_g^i \in \mathcal{A}(C)} \frac{Pr'}{\cancel{Pr}(a_g^i | x^i)} \right)$$

[Gimpel and Smith (2010)]

Learning

- Incorporate this loss with *softmax-margin* by adding it as a feature to the pairwise model

$$\sum_i \log \left(\sum_{a_g^i \in \mathcal{A}(C)} \frac{Pr'}{\cancel{Pr}(a_g^i | x^i)} \right) + \lambda \|w\|_1$$

[Gimpel and Smith (2010)]

Experiments

Experiments

- CoNLL 2011 dataset, system mentions from Lee et al. (2011)

Experiments

- CoNLL 2011 dataset, system mentions from Lee et al. (2011)
- Baselines:
  - Pairwise system
  - Centralized entity-level system following Rahman and Ng (2009)

Experiments

- CoNLL 2011 dataset, system mentions from Lee et al. (2011)
- Baselines:
  - Pairwise system
  - Centralized entity-level system following Rahman and Ng (2009)
- Two settings:
  - Synthetic features to contrast architectures
  - Standard entity features

Synthetic Properties

Synthetic Properties

- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

Synthetic Properties

- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

[James Reed]

[Reed]

[his]

Synthetic Properties

- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

[James Reed]
[Reed] $\longrightarrow$ 5
[his]

Synthetic Properties

- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

[James Reed]
[Reed]
[his]

$\longrightarrow$ 5

[Rose Brooks]
[Brooks]

$\longrightarrow$ 3

Synthetic Properties

- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

[James Reed]
[Reed]
[his]
$\rightarrow$ 5

[Rose Brooks]
[Brooks]
$\rightarrow$ 3

[New York]
$\rightarrow$ 5

Synthetic Properties

- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

[James Reed]
[Reed]
[his]
$\rightarrow$ 5

[Rose Brooks]
[Brooks]
$\rightarrow$ 3

[New York]
$\rightarrow$ 5

- For each mention, sample from Dirichlet peaked on ℓ

Synthetic Properties

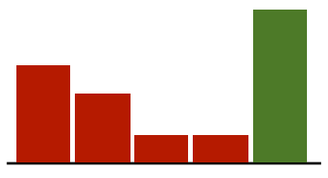
- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

[James Reed]
[Reed]
[his]
$\rightarrow$ 5

[Rose Brooks]
[Brooks]
$\rightarrow$ 3

[New York] $\rightarrow$ 5

- For each mention, sample from Dirichlet peaked on ℓ

James
Reed
$\rightarrow$ 

Synthetic Properties

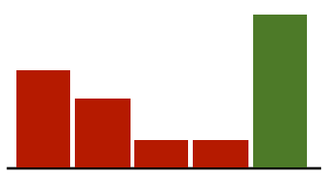
- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

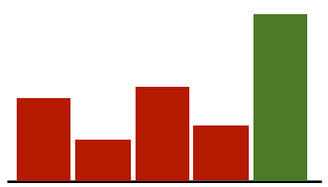
[James Reed]
[Reed]
[his]
$\rightarrow$ 5

[Rose Brooks]
[Brooks]
$\rightarrow$ 3

[New York] $\rightarrow$ 5

- For each mention, sample from Dirichlet peaked on ℓ

James Reed $\rightarrow$ 

Reed $\rightarrow$ 

Synthetic Properties

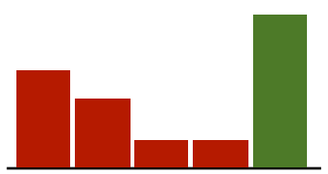
- For each *gold* cluster, label $\ell \sim U[\{1, 2, 3, 4, 5\}]$

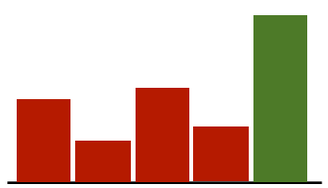
[James Reed]
[Reed]
[his]
$\rightarrow$ 5

[Rose Brooks]
[Brooks]
$\rightarrow$ 3

[New York] $\rightarrow$ 5

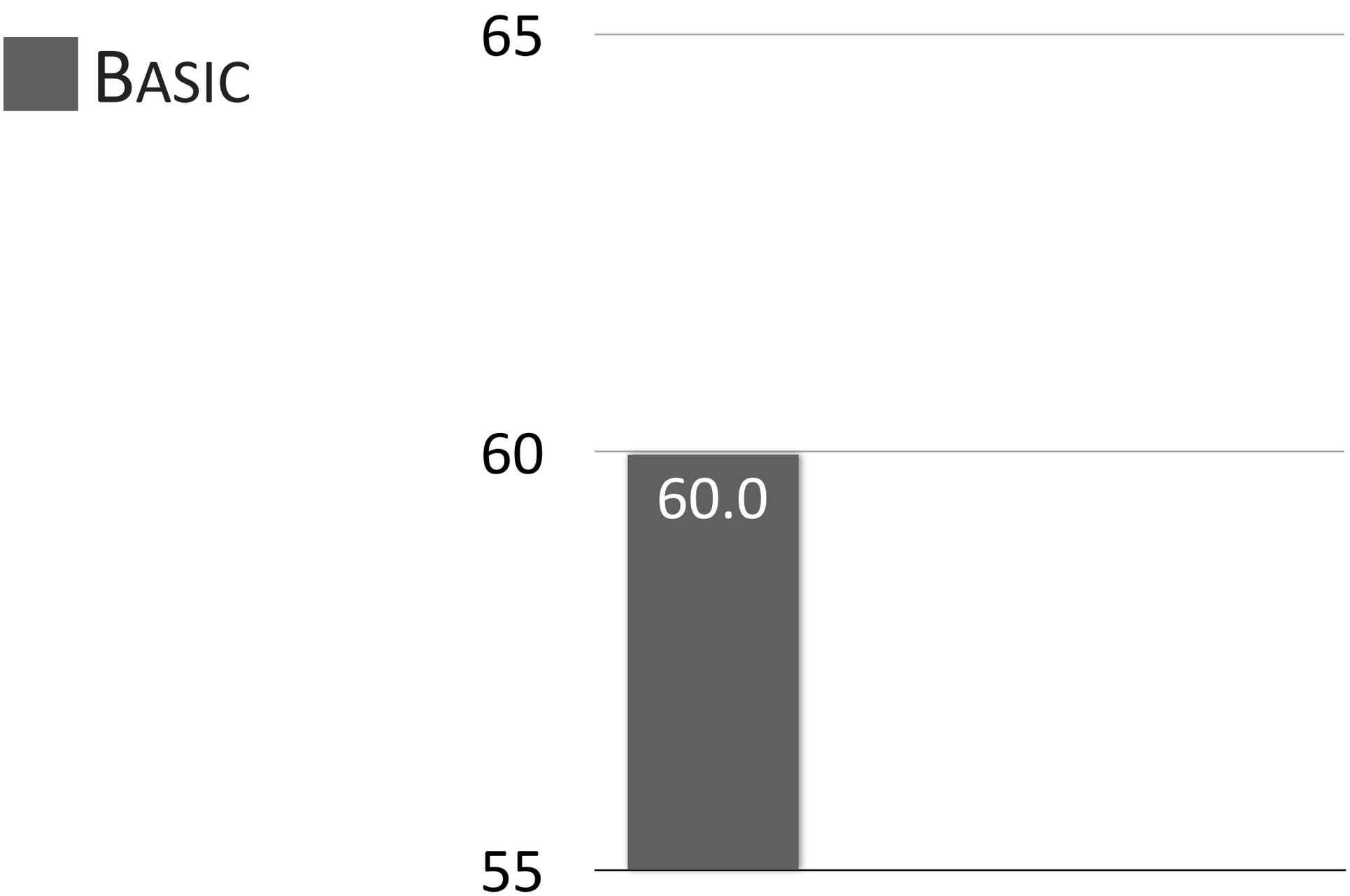
- For each mention, sample from Dirichlet peaked on ℓ

James Reed $\rightarrow$ 

Reed $\rightarrow$ 

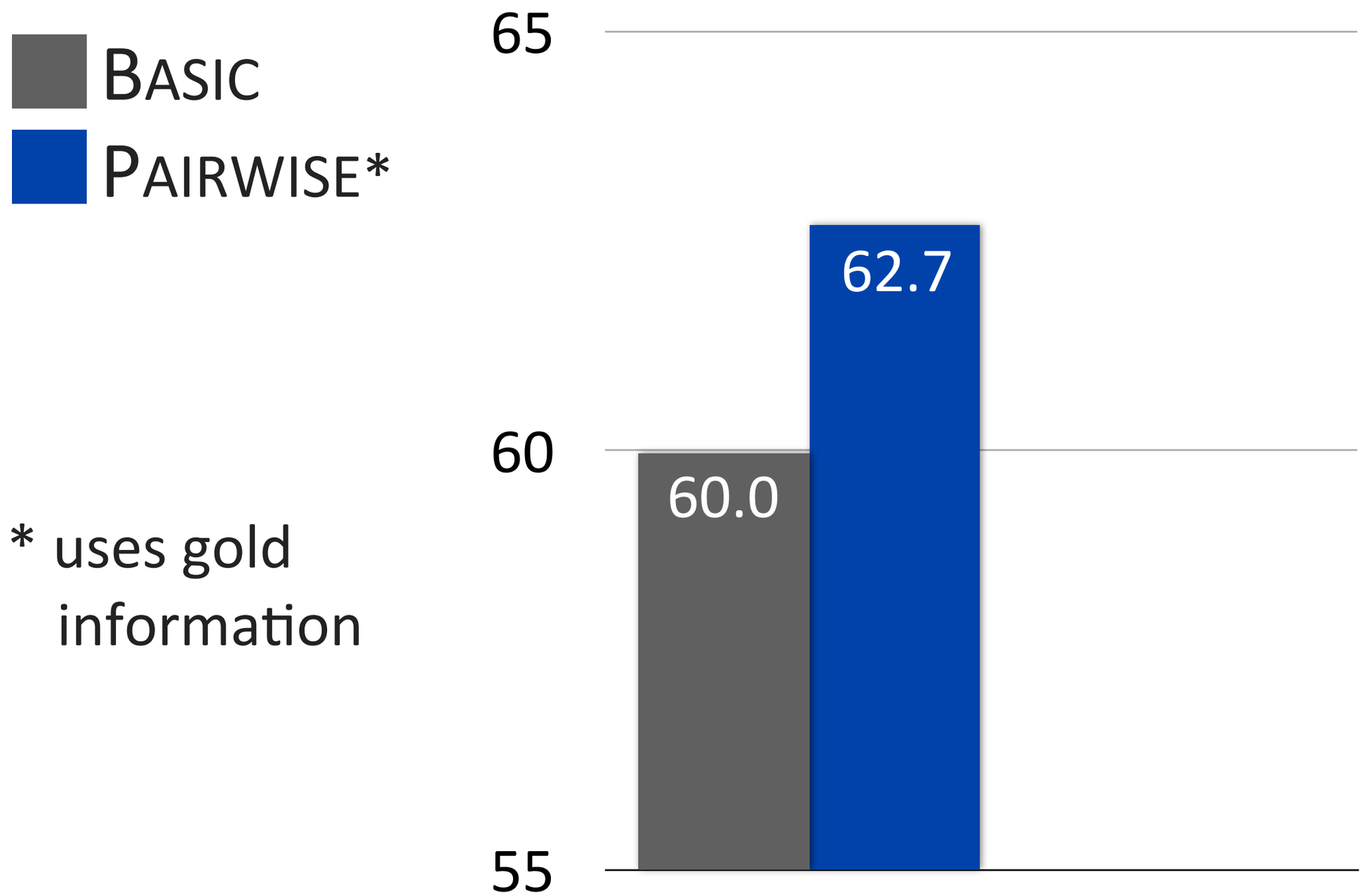
his $\rightarrow$ 

Synthetic Properties



(CoNLL scores, 10-fold cross-validation on train set)

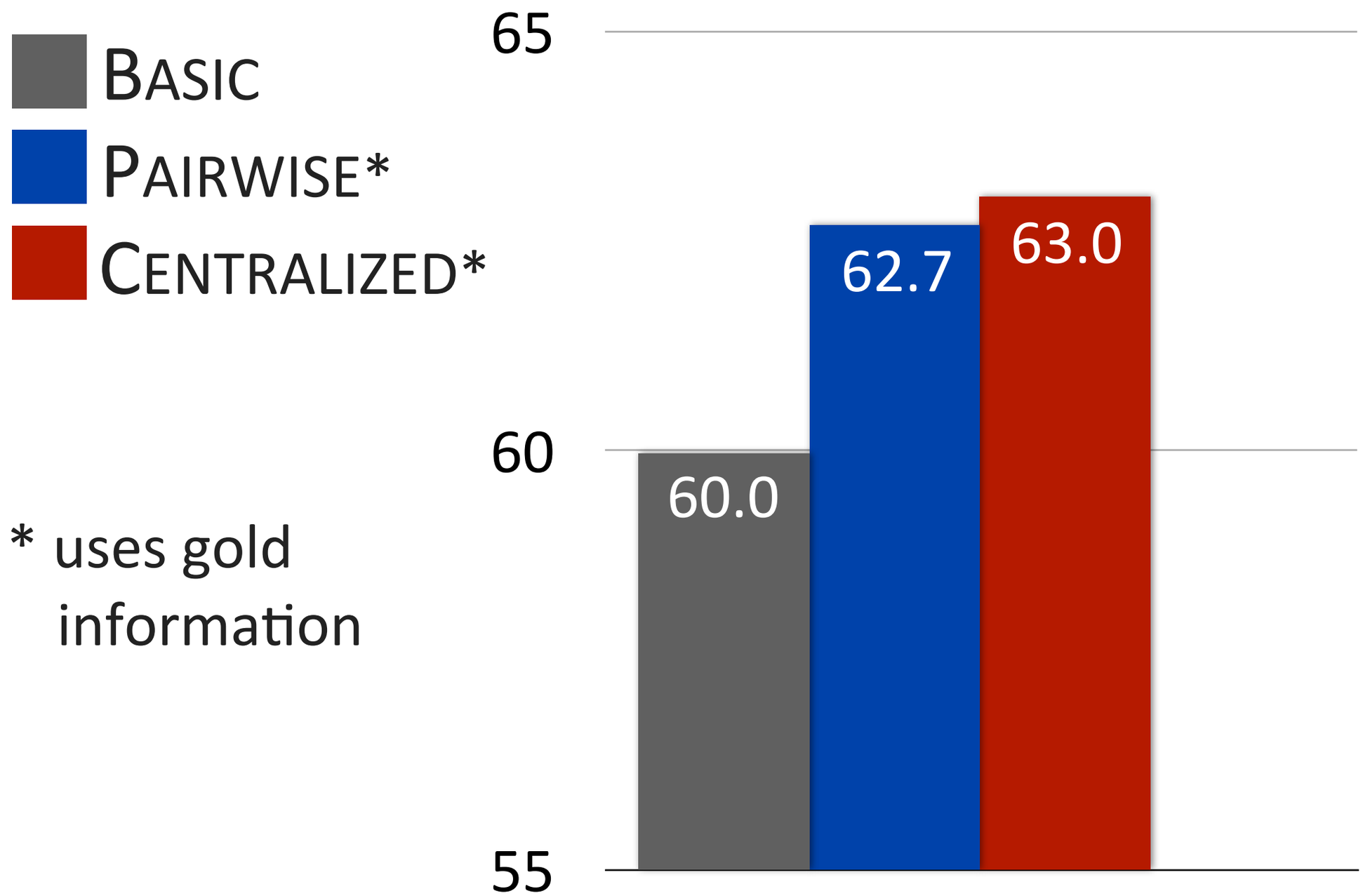
Synthetic Properties



* uses gold information

(CoNLL scores, 10-fold cross-validation on train set)

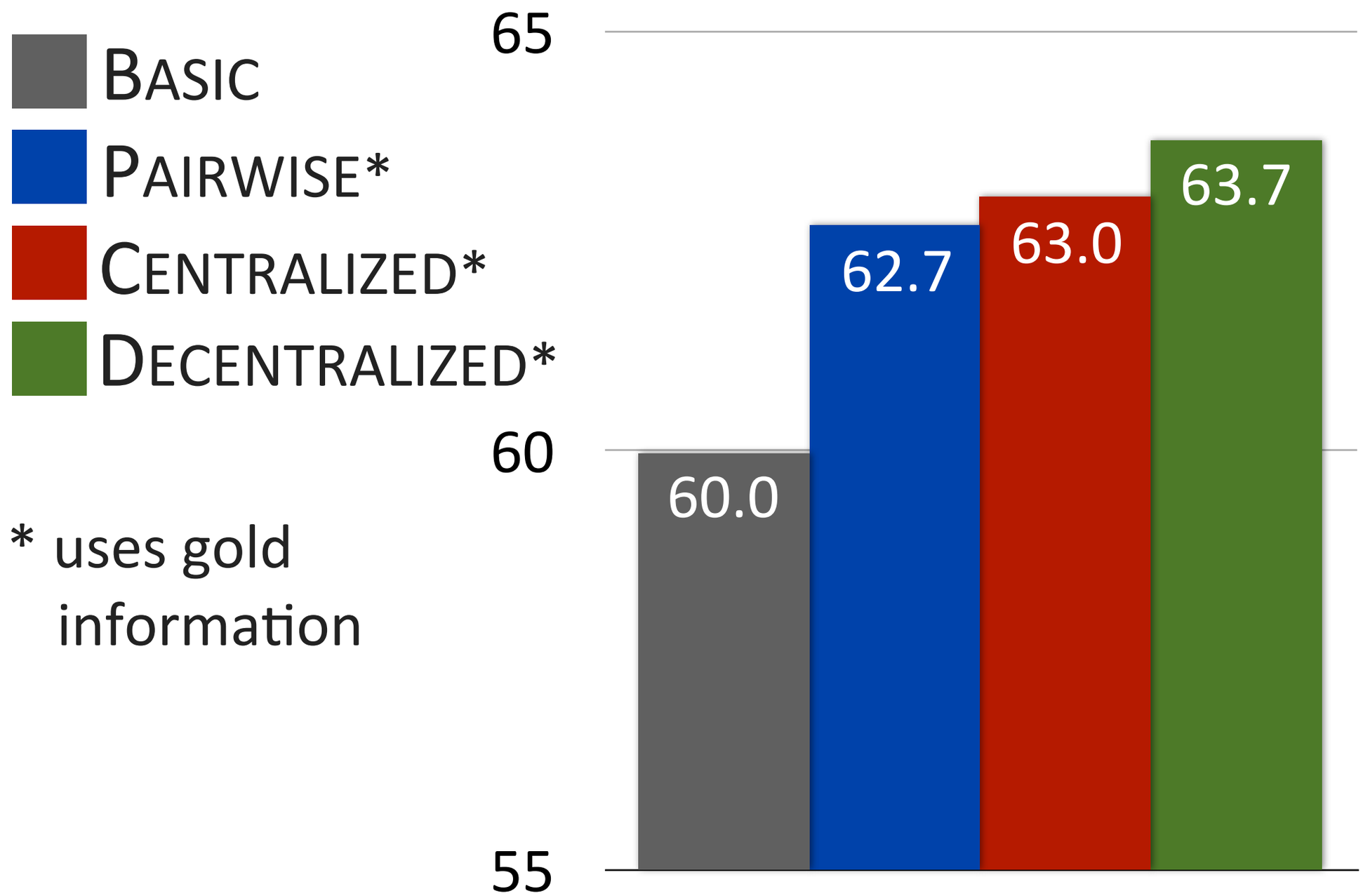
Synthetic Properties



* uses gold information

(CoNLL scores, 10-fold cross-validation on train set)

Synthetic Properties



* uses gold information

(CoNLL scores, 10-fold cross-validation on train set)

ϕ-feature Properties

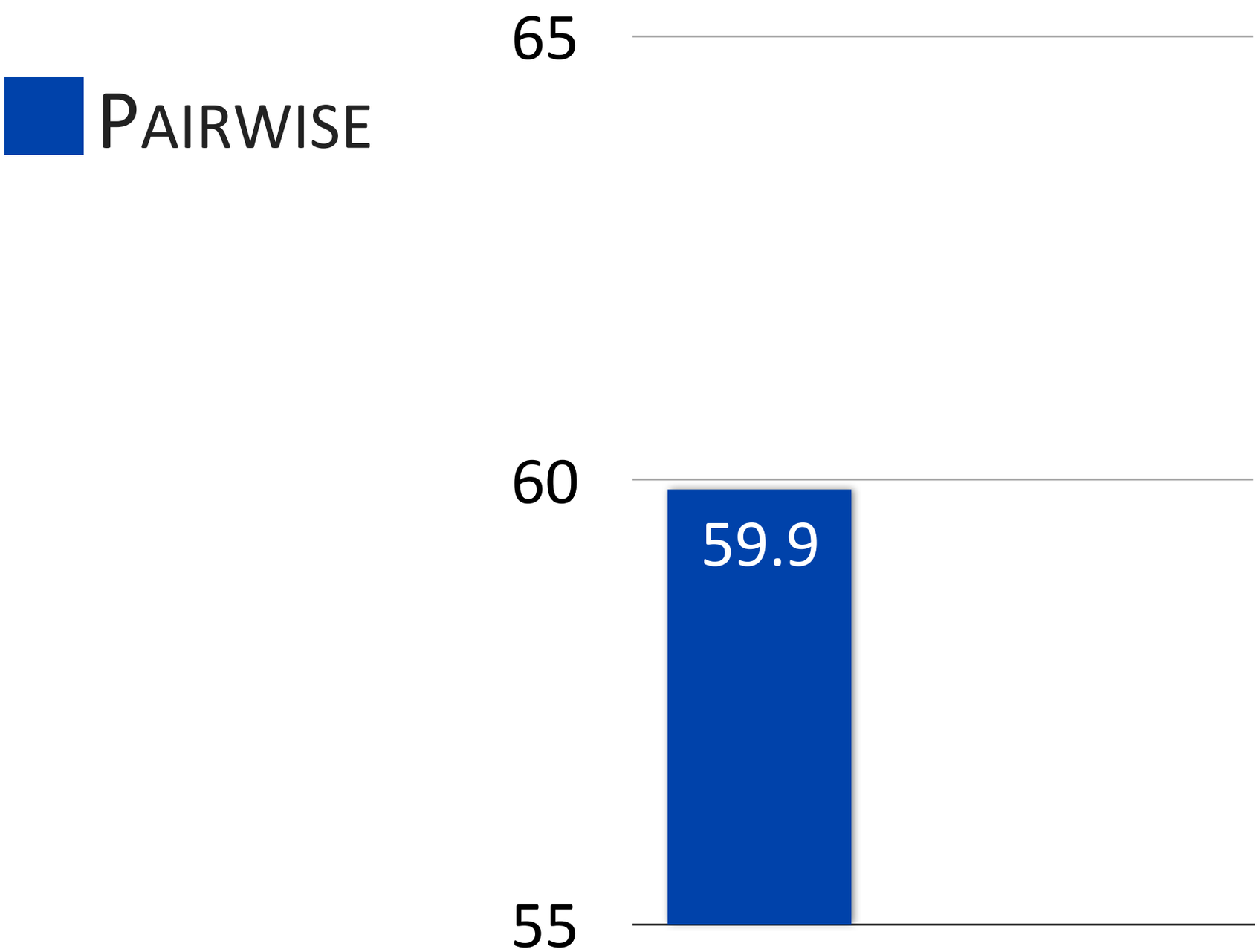
[Rahman and Ng (2009), Lee et al. (2011), inter alia]

ϕ-feature Properties

- Properties based on linguistic ϕ-features:
  - Number
  - Gender
  - Animacy
  - NE type

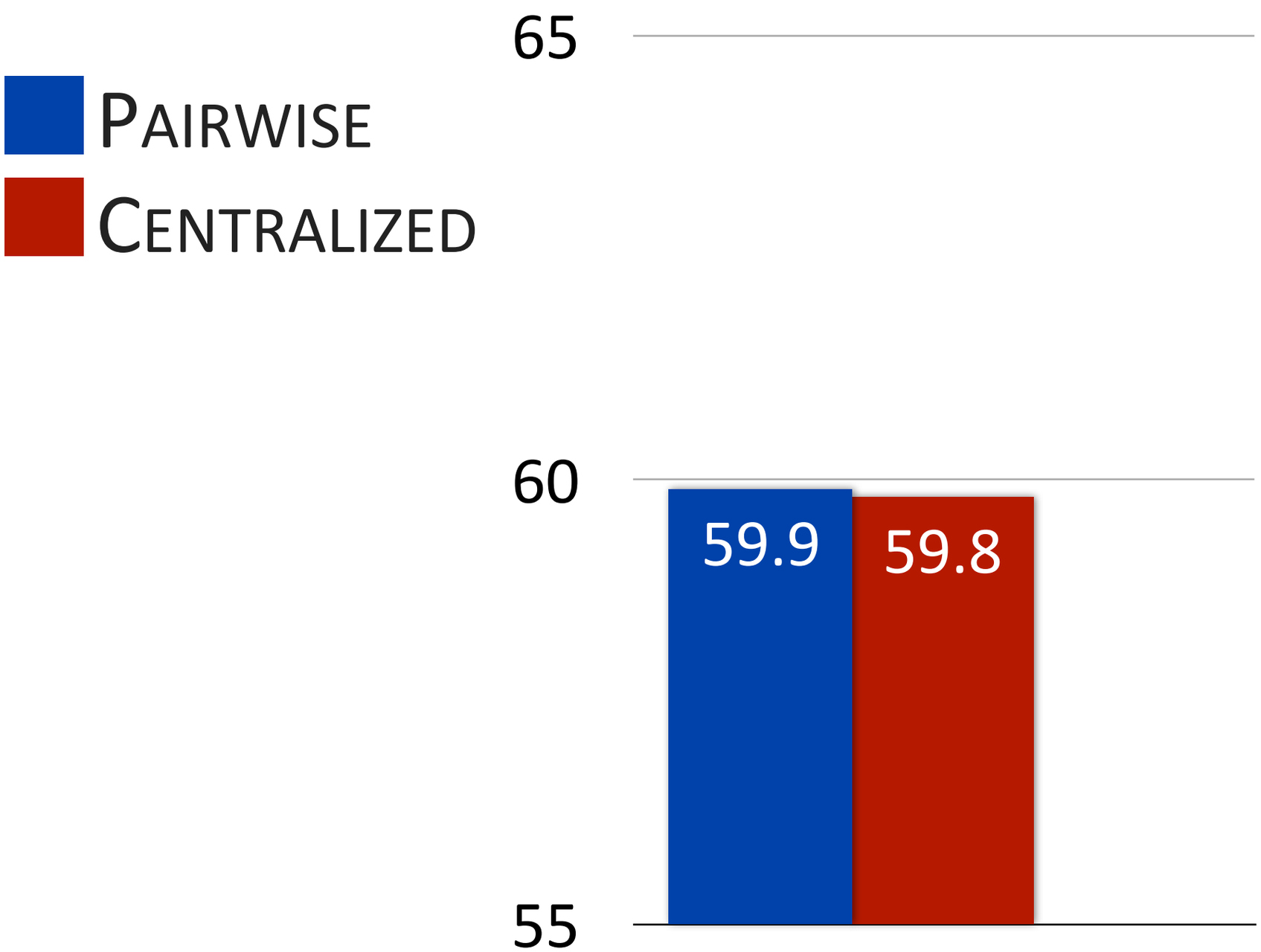
[Rahman and Ng (2009), Lee et al. (2011), inter alia]

ϕ-feature Properties



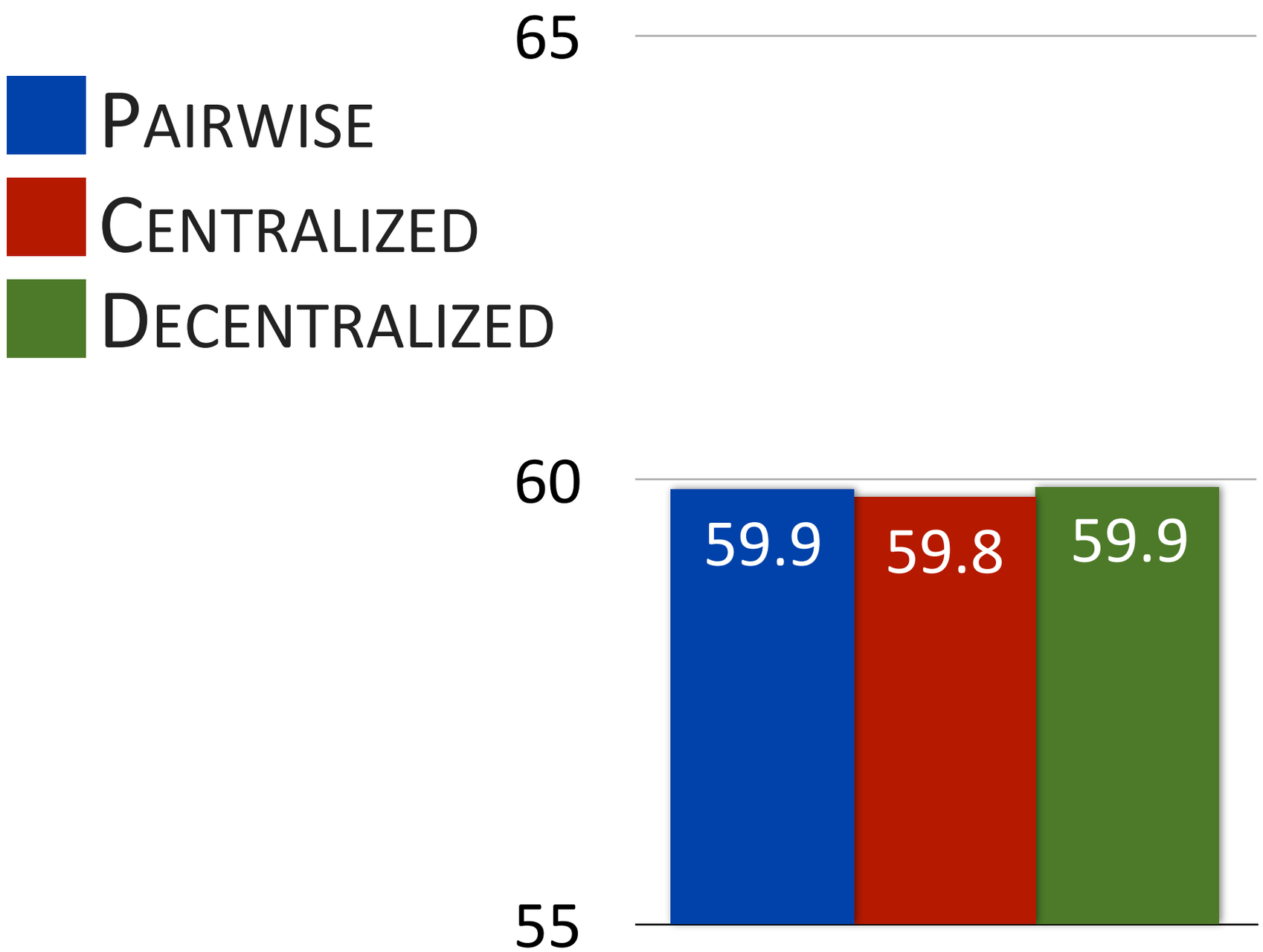
(CoNLL scores, 10-fold cross-validation on train set)

ϕ-feature Properties



(CoNLL scores, 10-fold cross-validation on train set)

ϕ-feature Properties



(CoNLL scores, 10-fold cross-validation on train set)

Semantic Properties

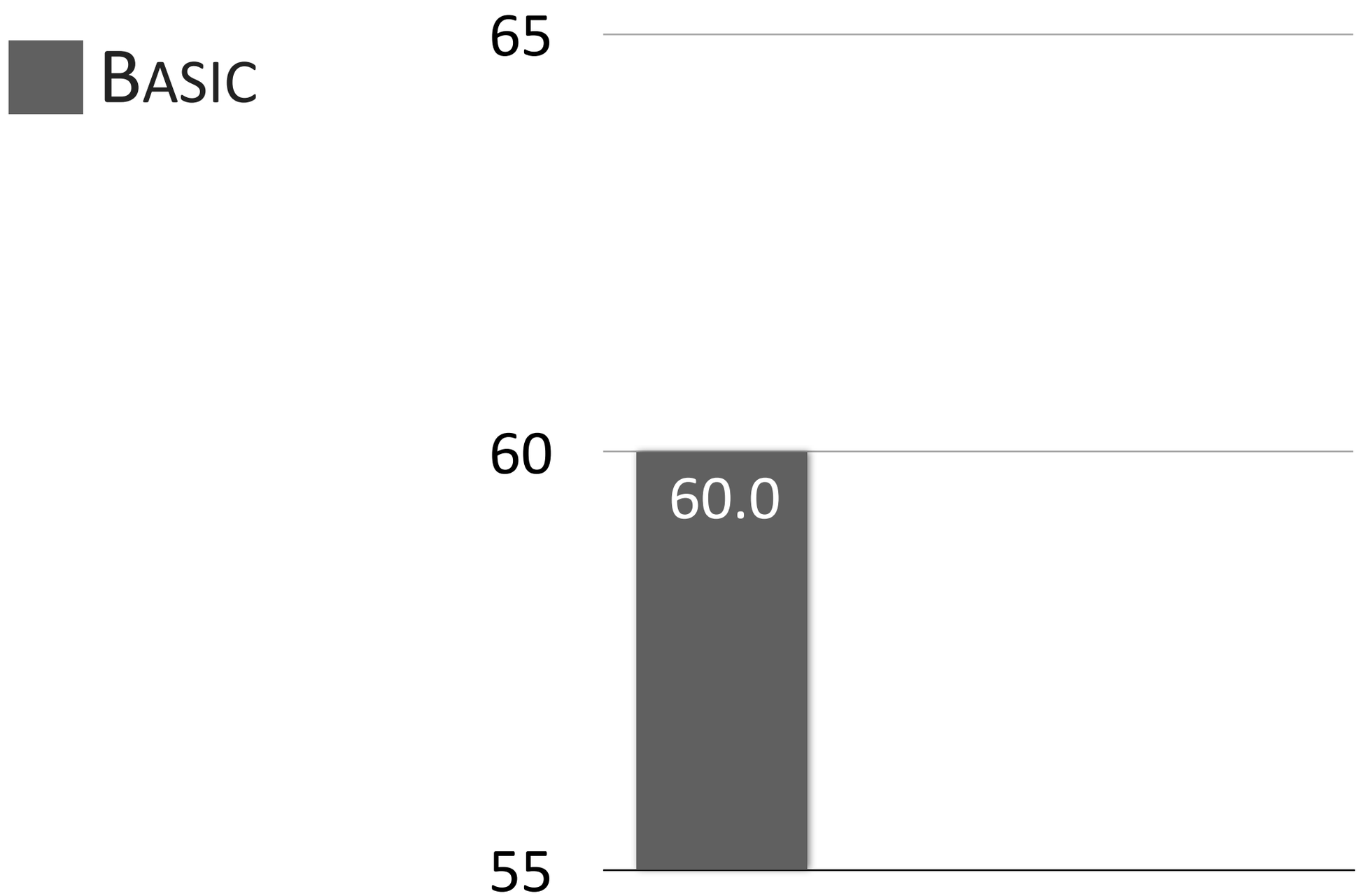
- ϕ-features do not capture fine-grained semantic distinctions between entities

Semantic Properties

- ϕ-features do not capture fine-grained semantic distinctions between entities

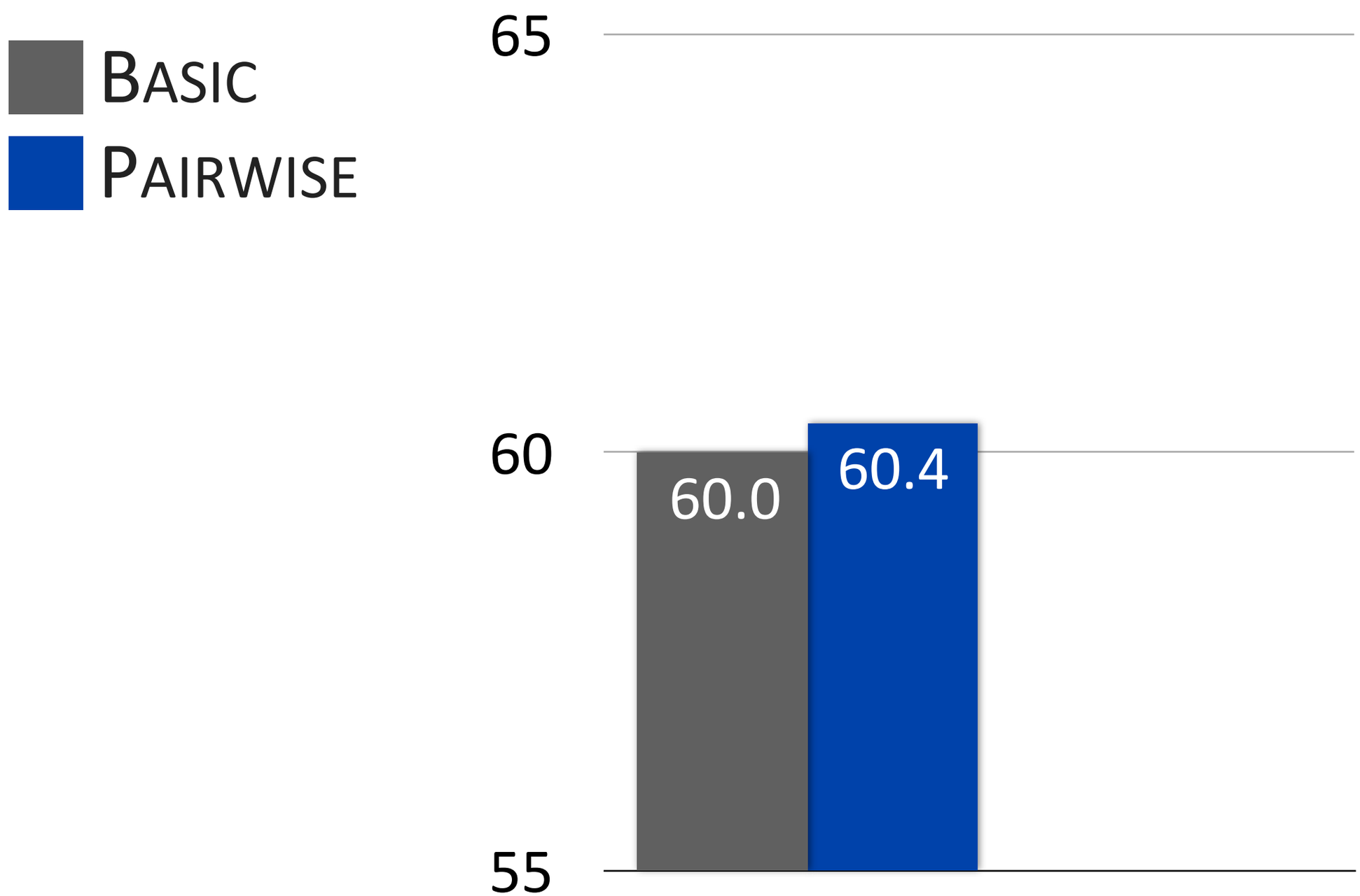
- Use properties derived from unsupervised clustering of headwords and their governors

Semantic Properties



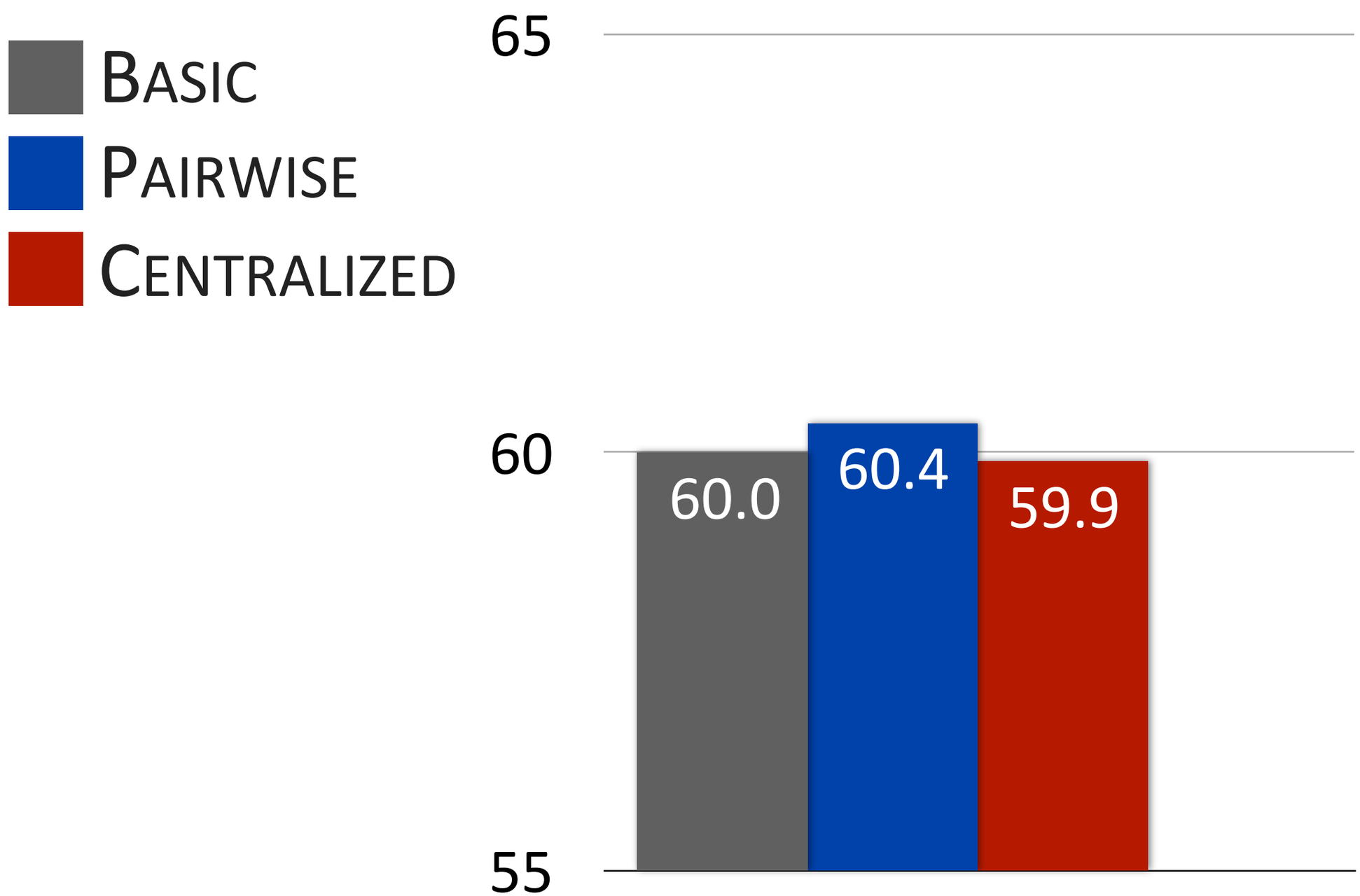
(CoNLL scores, 10-fold cross-validation on train set)

Semantic Properties



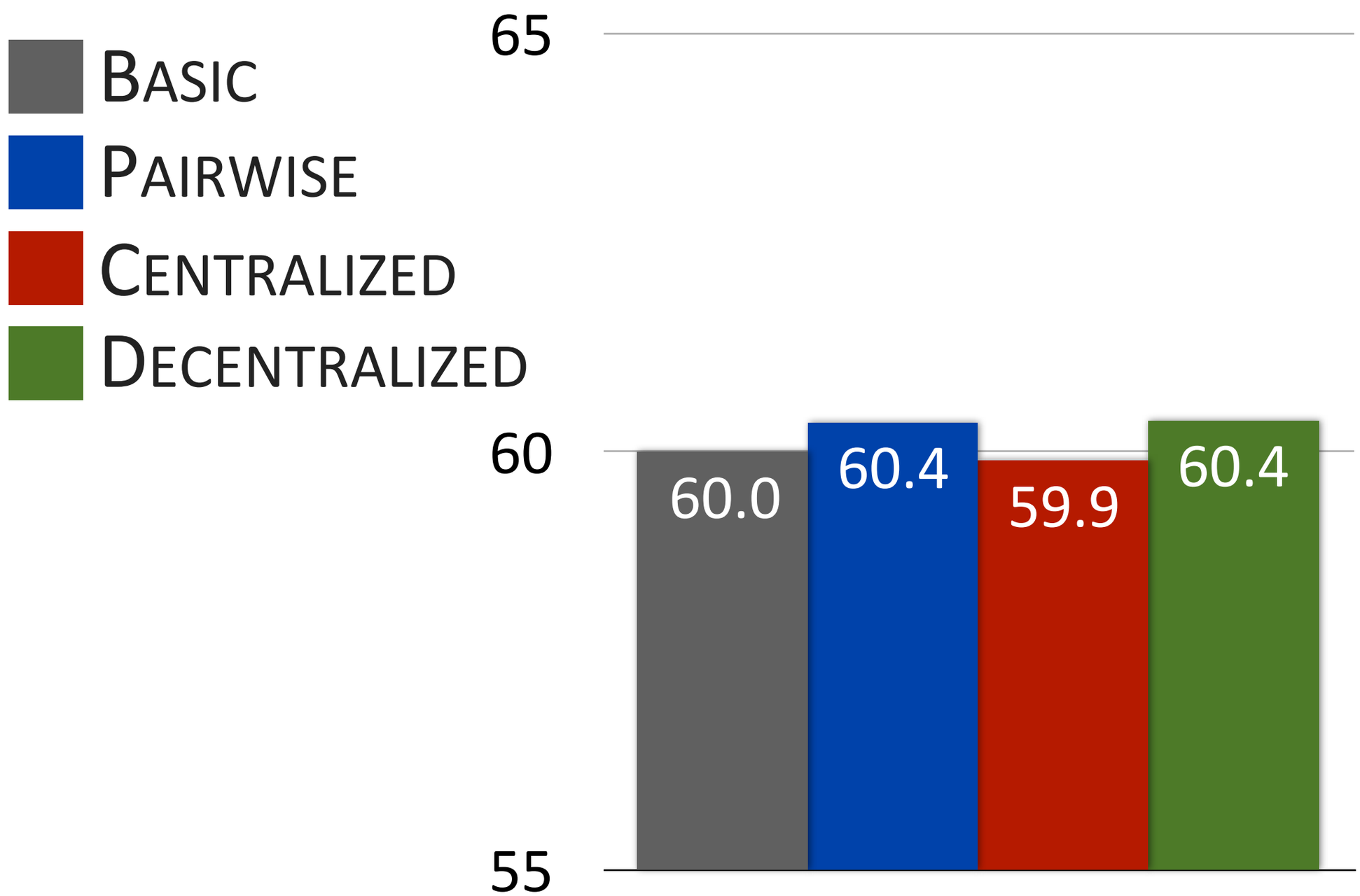
(CoNLL scores, 10-fold cross-validation on train set)

Semantic Properties



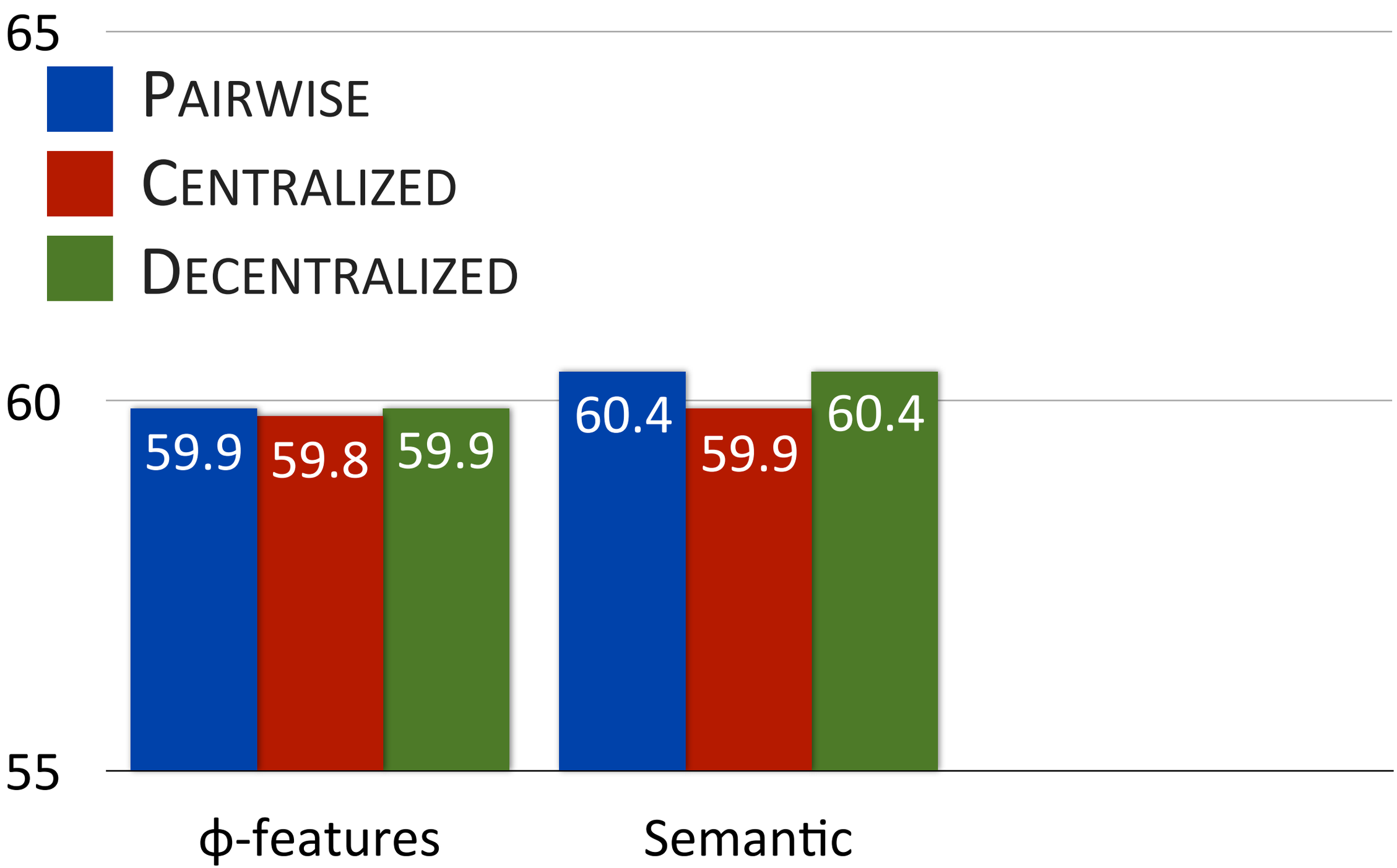
(CoNLL scores, 10-fold cross-validation on train set)

Semantic Properties

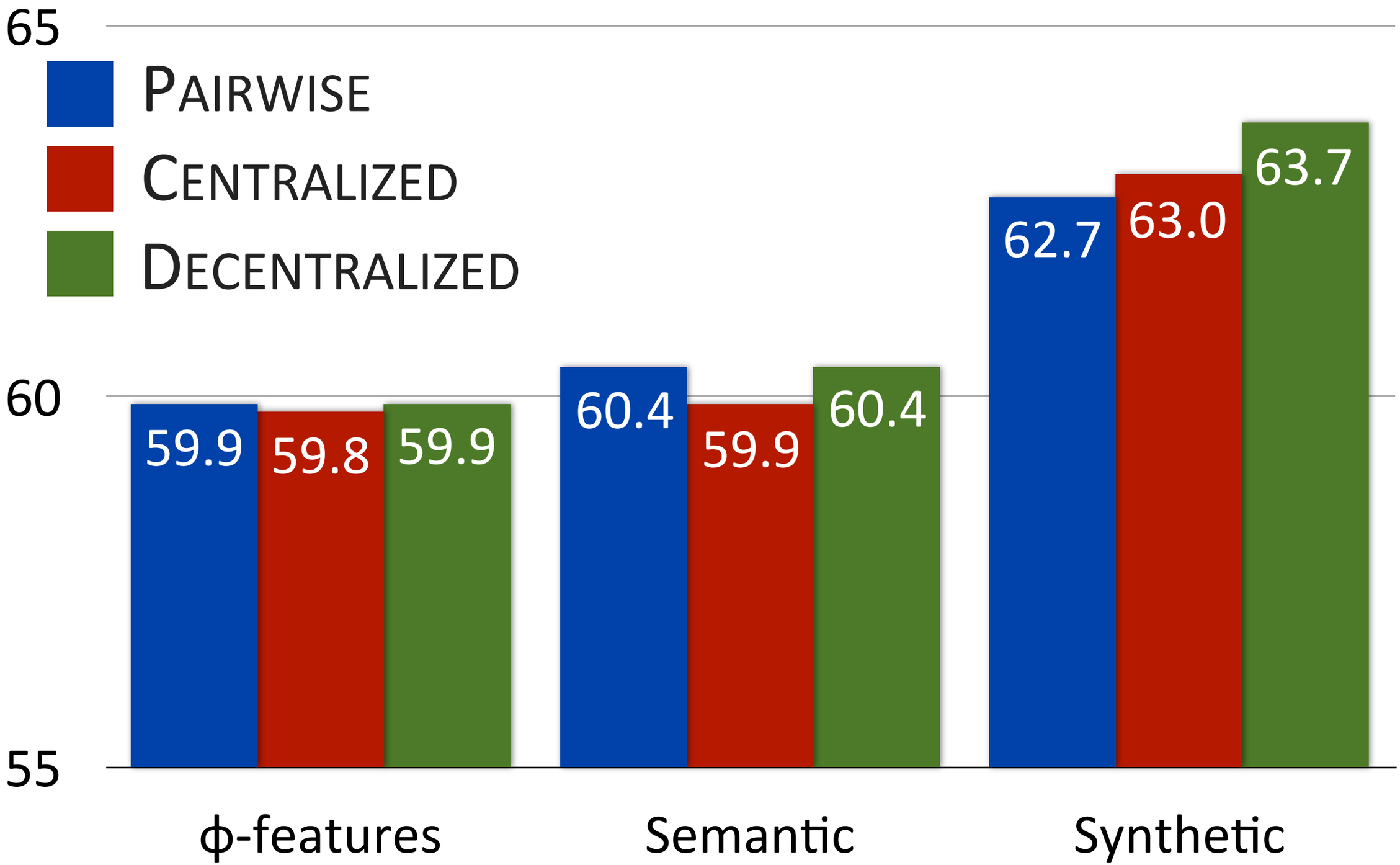


(CoNLL scores, 10-fold cross-validation on train set)

Overall Results



Overall Results



Conclusion

- Our model effectively integrates entity-level features in an end-to-end way

Conclusion

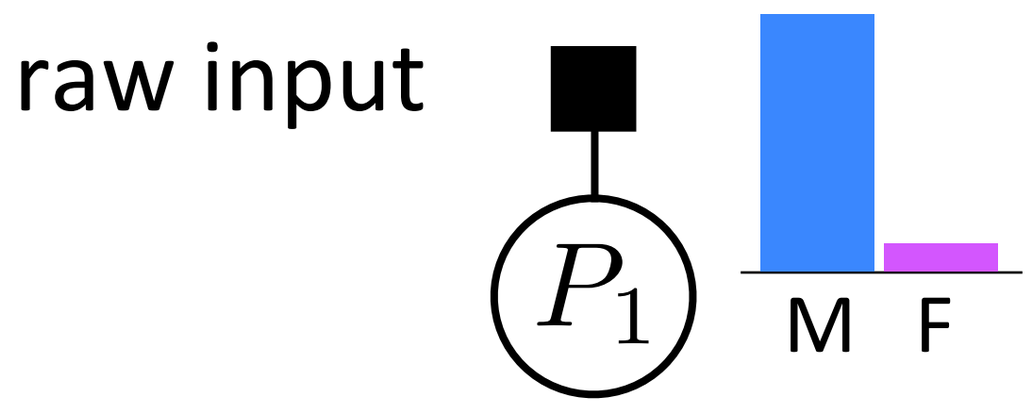
- Our model effectively integrates entity-level features in an end-to-end way
- Good entity-level features are hard to find: simple ϕ-feature and semantic type propagation give little benefit

Conclusion

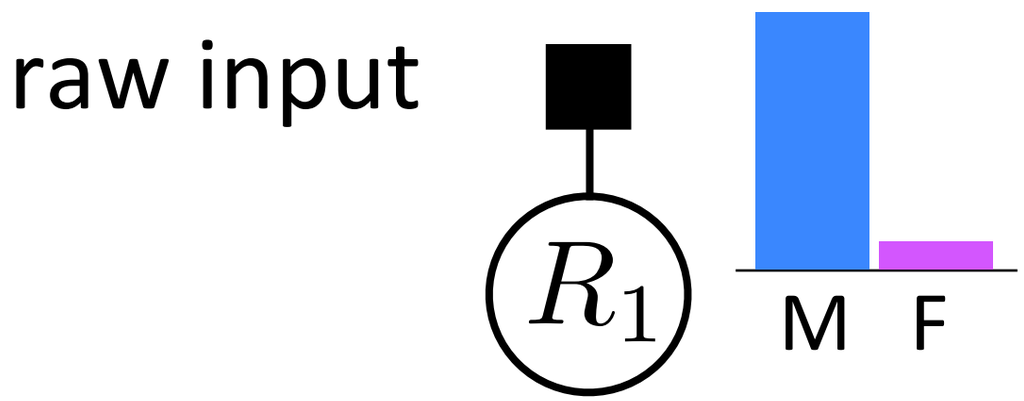
- Our model effectively integrates entity-level features in an end-to-end way
- Good entity-level features are hard to find: simple ϕ-feature and semantic type propagation give little benefit

Thank you!

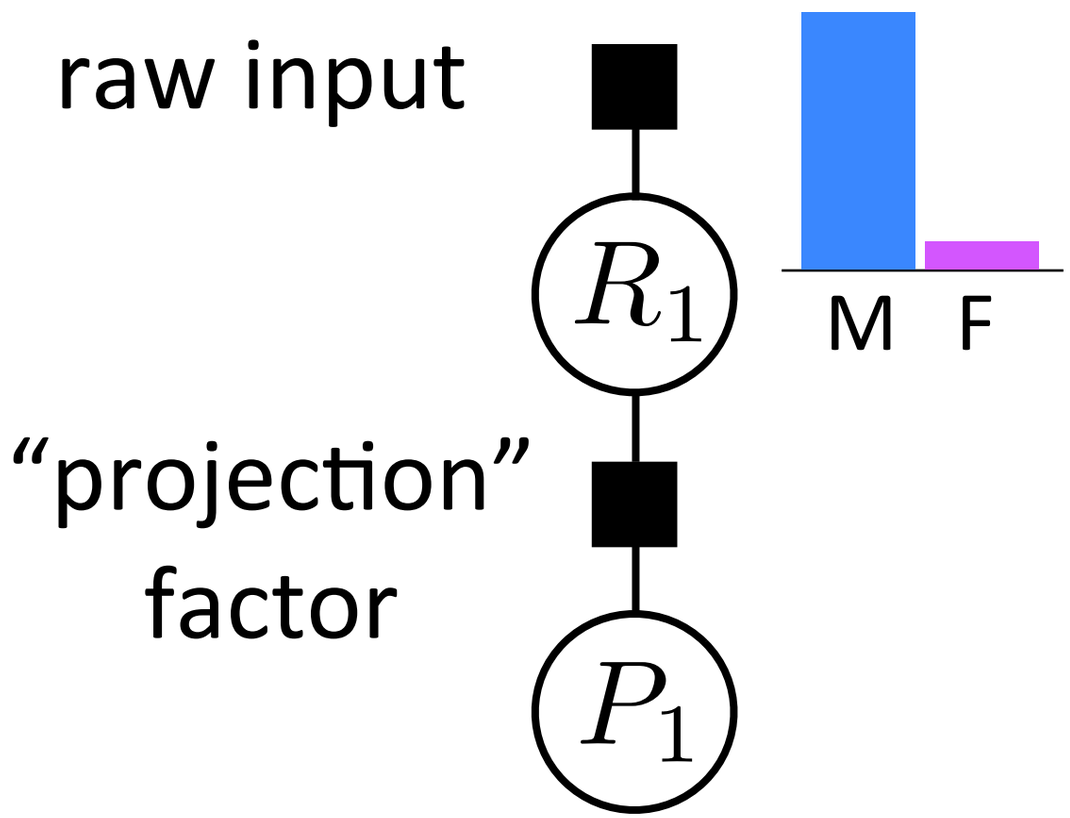
Projected Properties



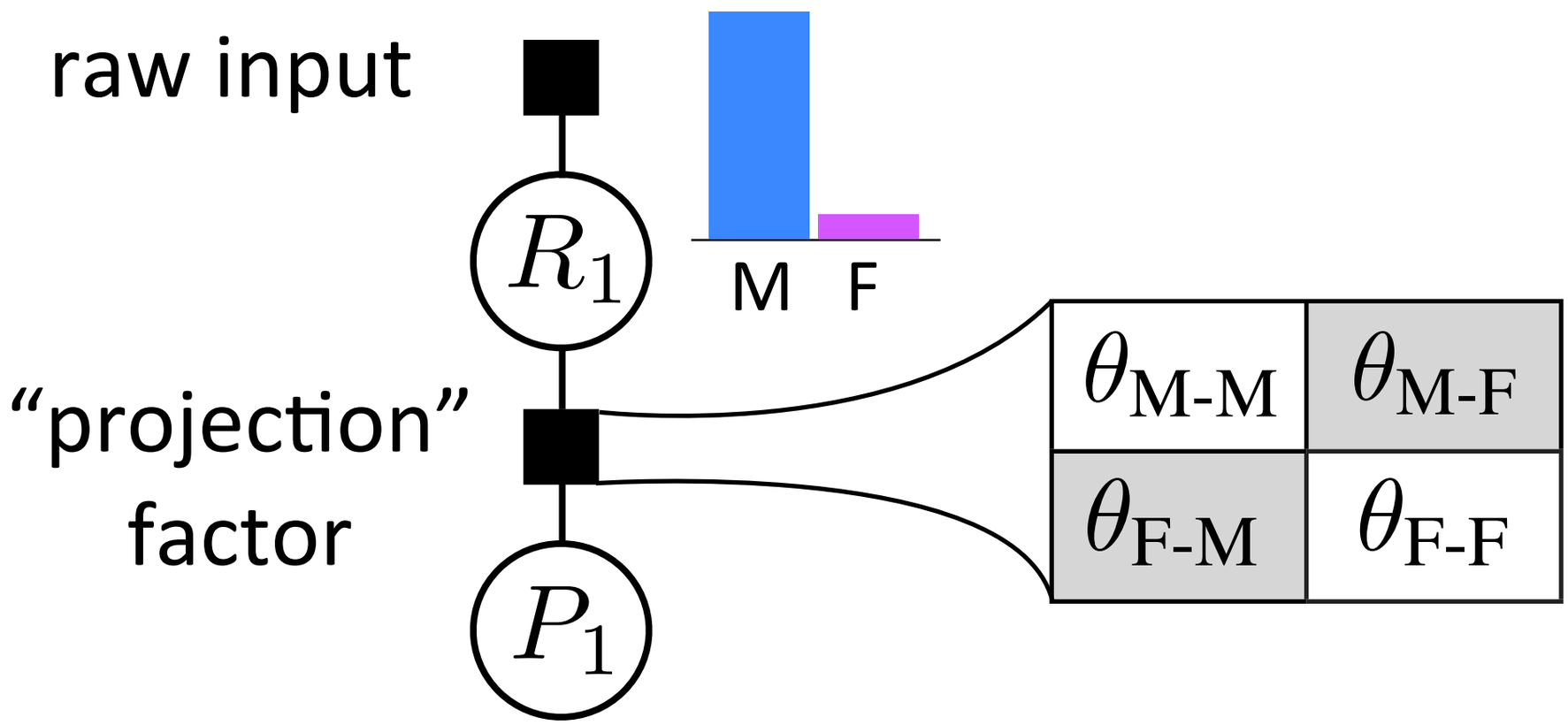
Projected Properties



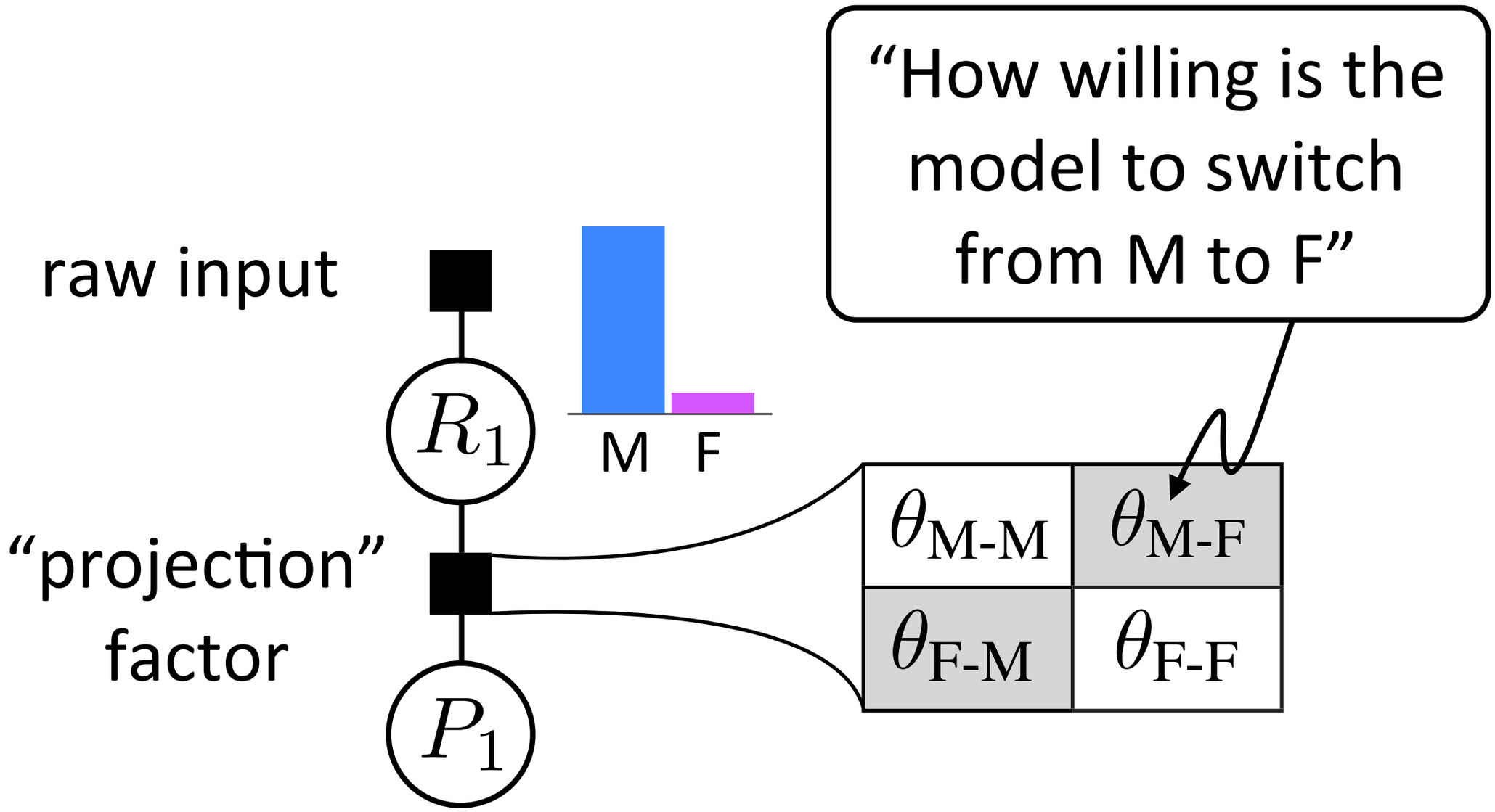
Projected Properties



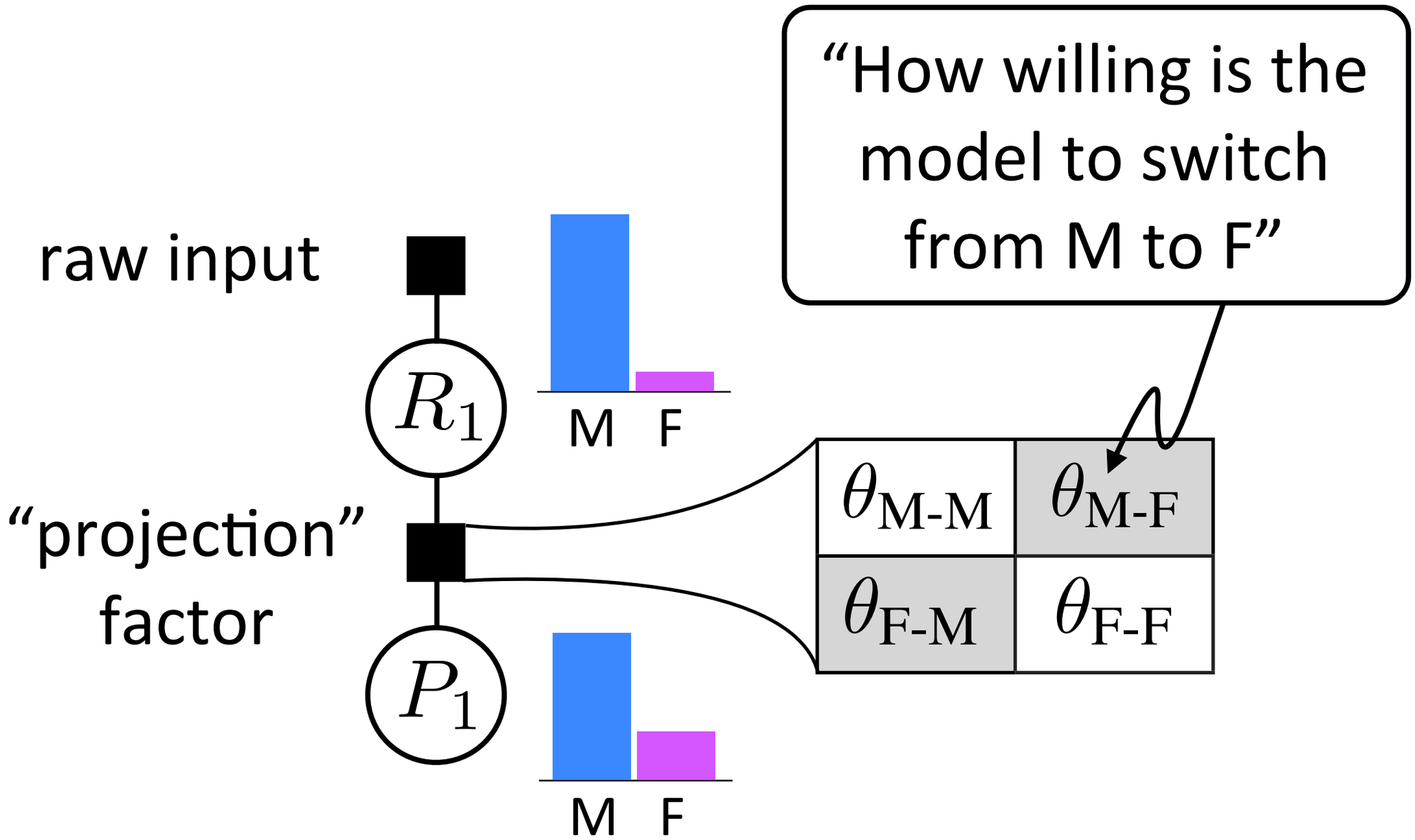
Projected Properties



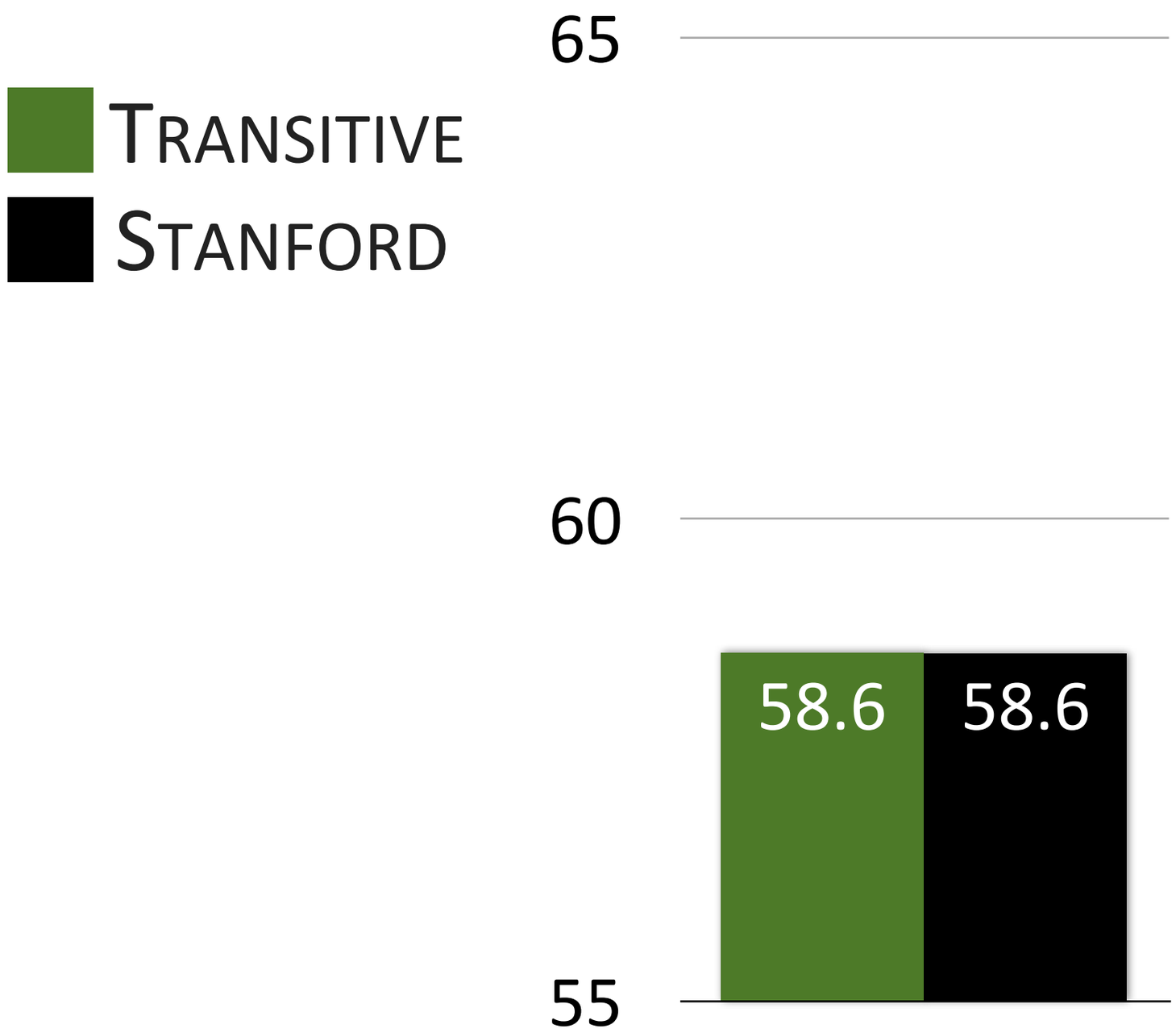
Projected Properties



Projected Properties



Final Results



(CoNLL scores, blind test set)