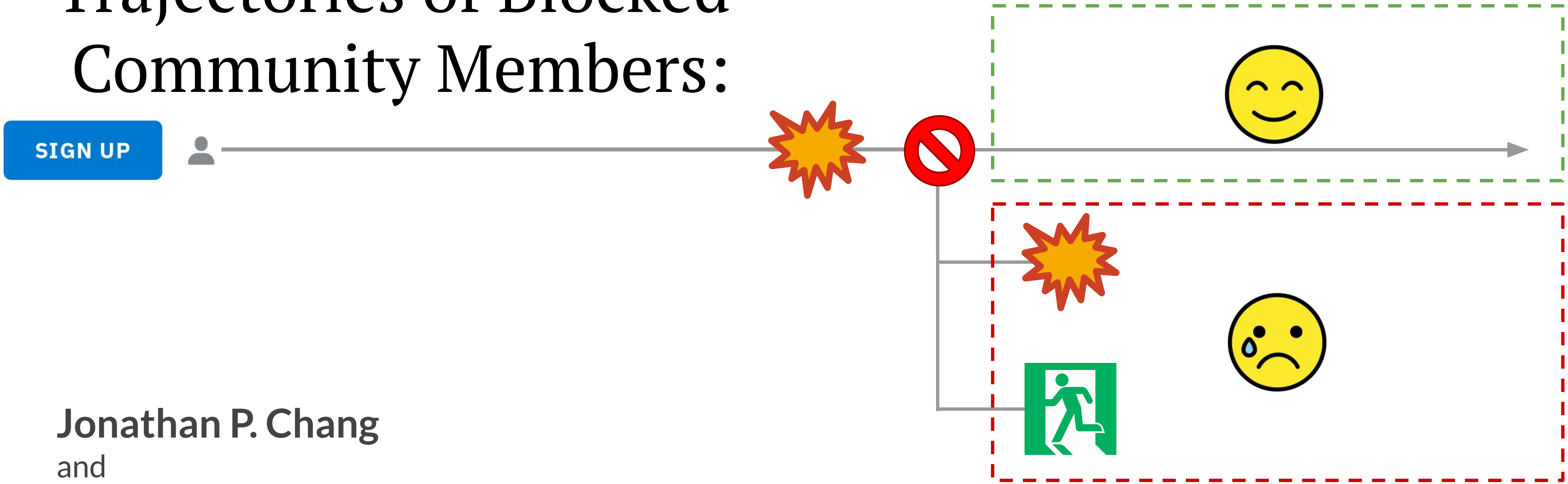
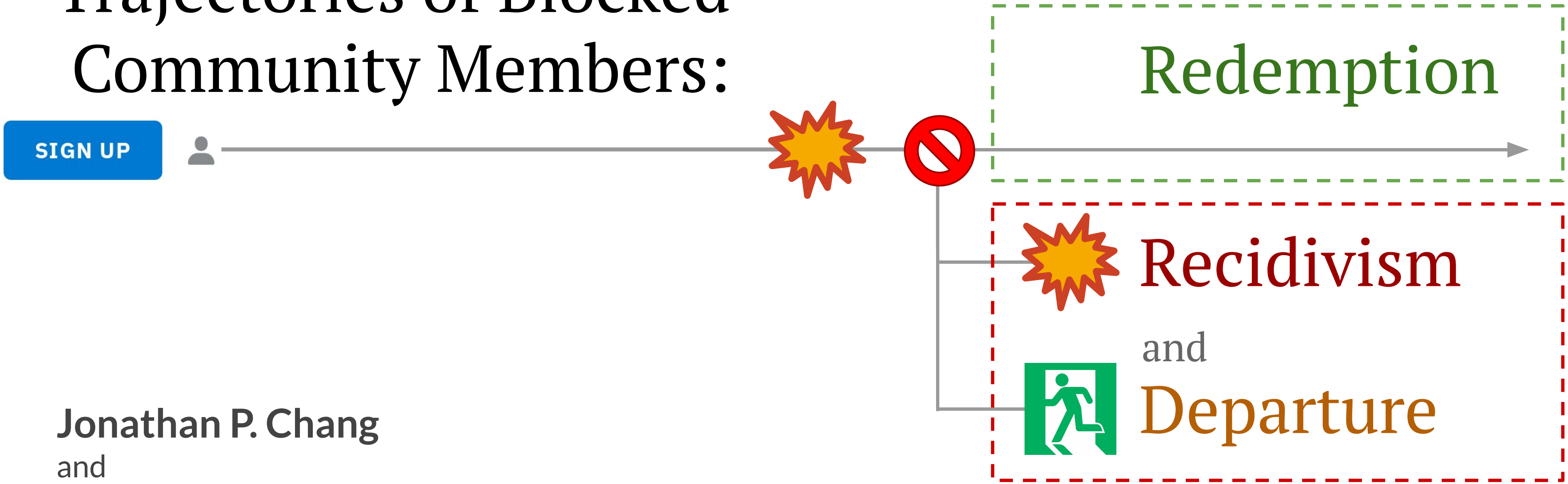
Trajectories of Blocked Community Members:

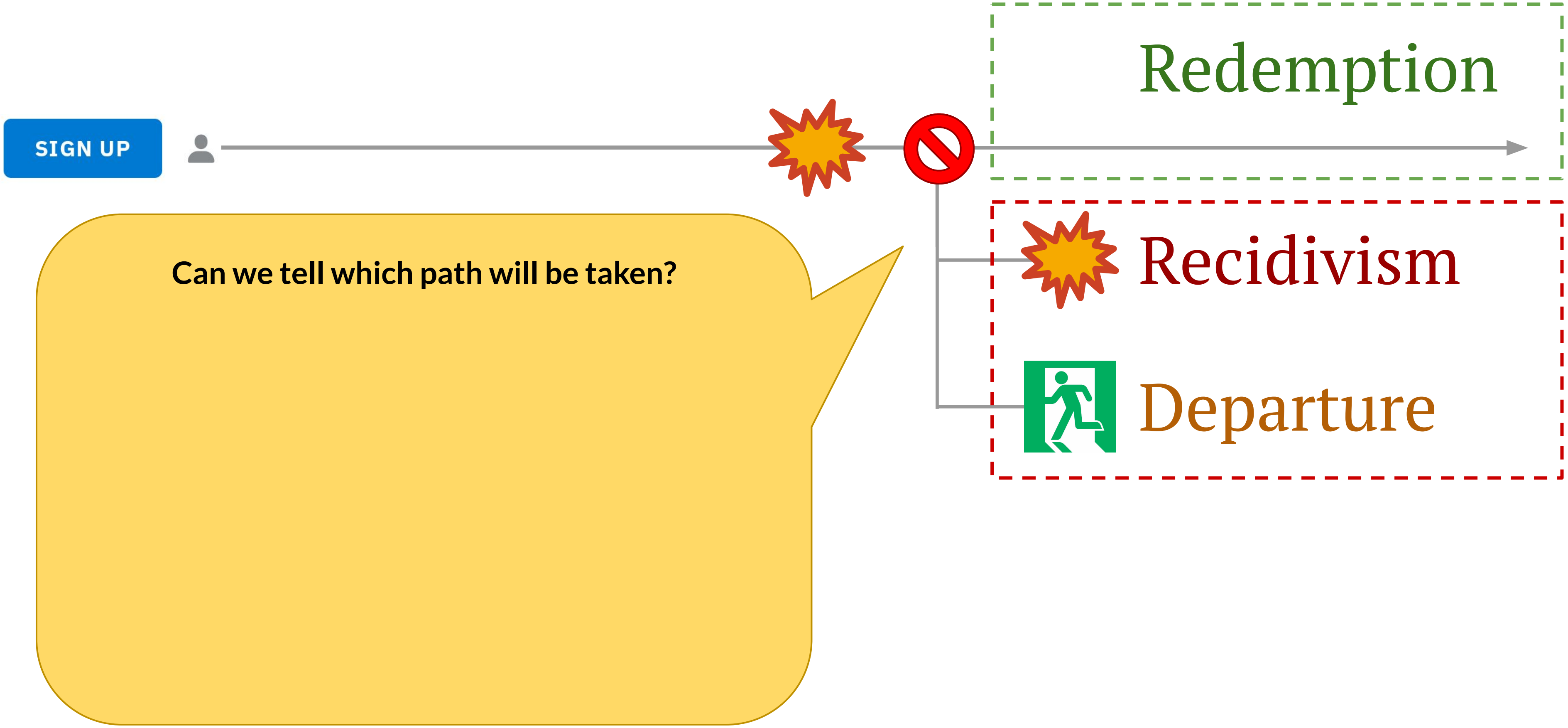


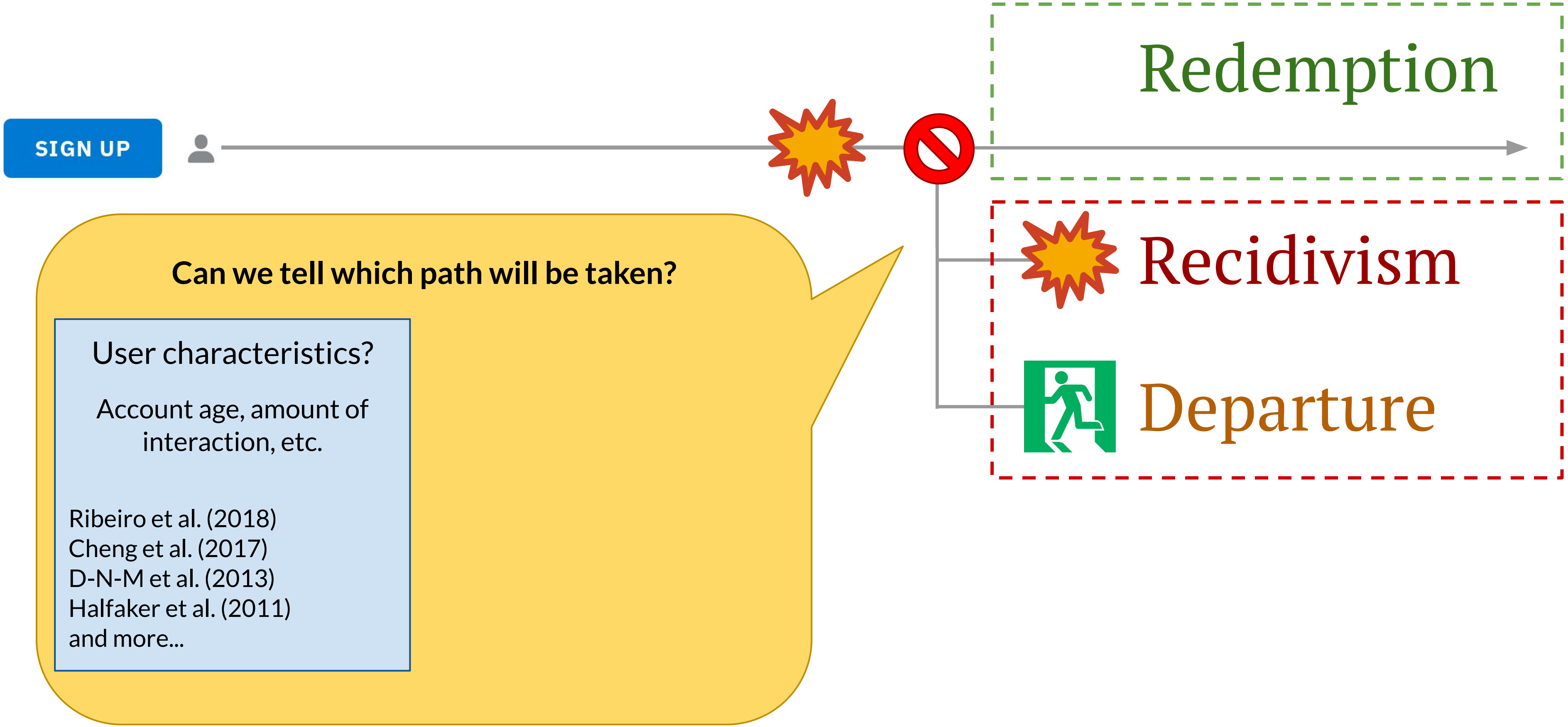
Jonathan P. Chang
and
Cristian Danescu-Niculescu-Mizil
Cornell University

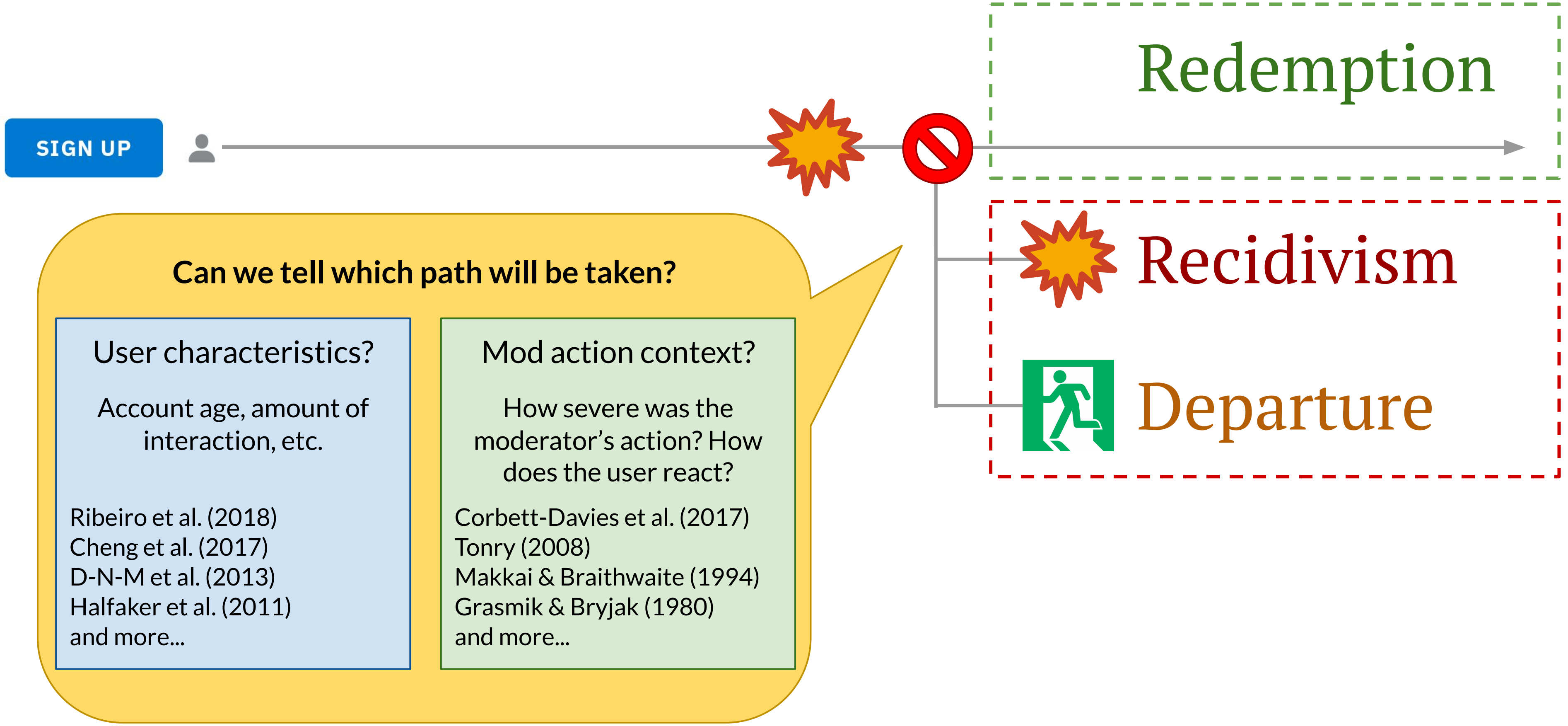
Trajectories of Blocked Community Members:

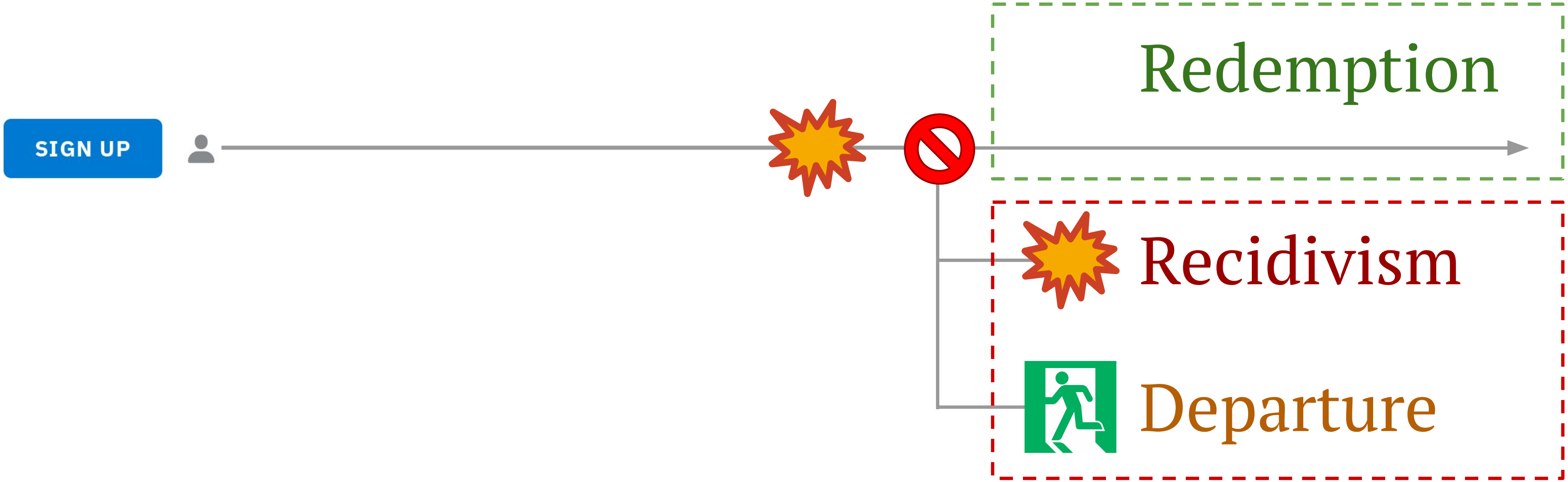


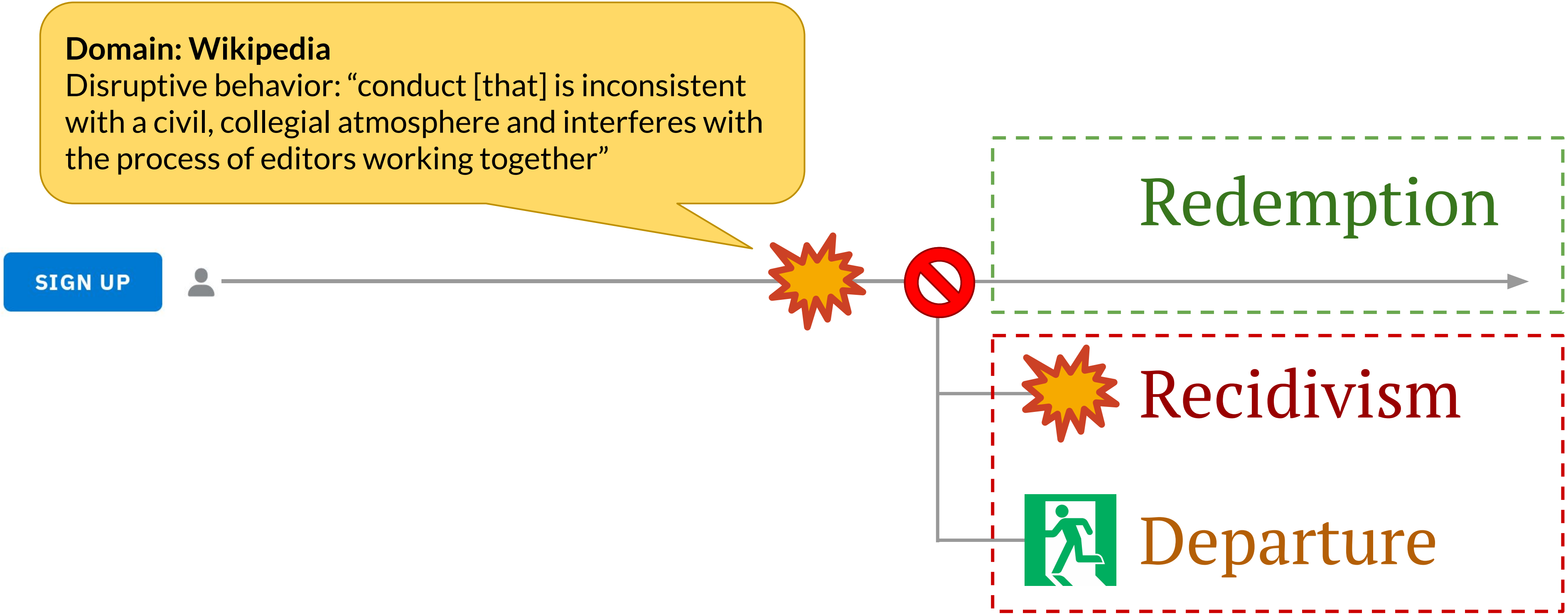
Jonathan P. Chang
and
Cristian Danescu-Niculescu-Mizil
Cornell University

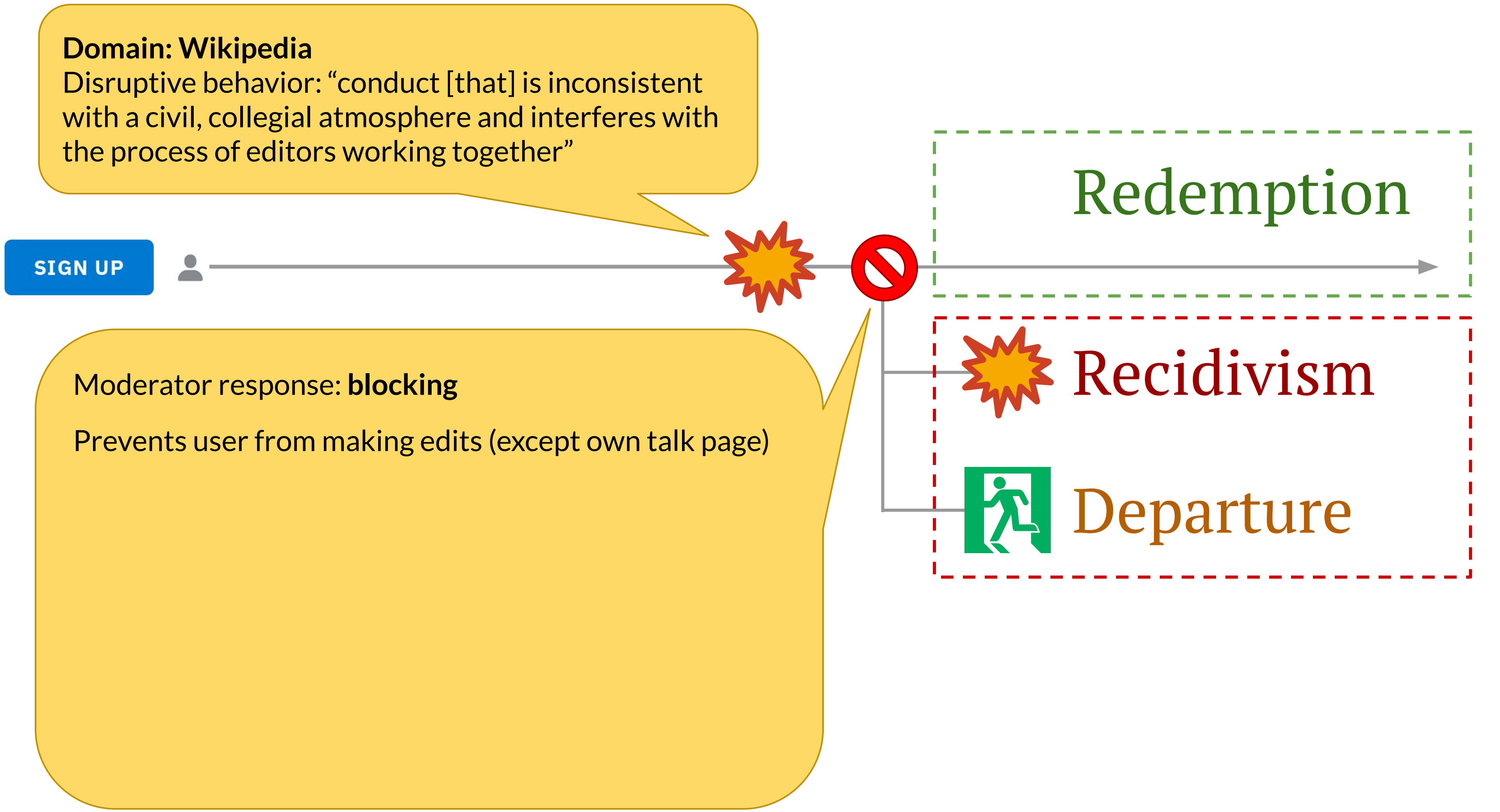

SIGN UP
Redemption
Recidivism
Departure
Can we tell which path will be taken?


SIGN UP
Redemption
Recidivism
Departure
Can we tell which path will be taken?
User characteristics?
Account age, amount of interaction, etc.
Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...

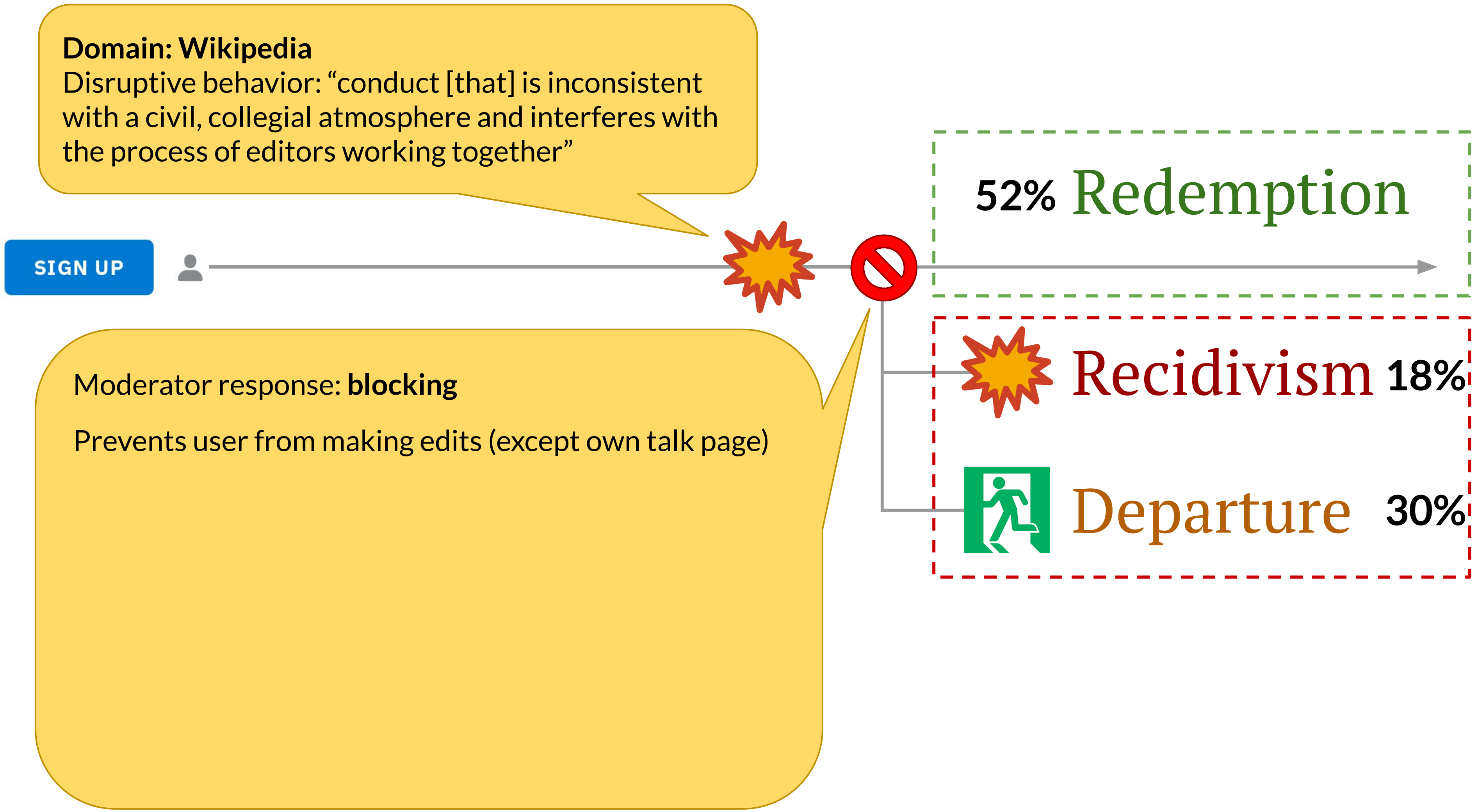

SIGN UP
Redemption
Recidivism
Departure
Can we tell which path will be taken?
User characteristics?
Account age, amount of interaction, etc.
Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...
Mod action context?
How severe was the moderator's action? How does the user react?
Corbett-Davies et al. (2017)
Tonry (2008)
Makkai & Braithwaite (1994)
Grasmik & Bryjak (1980)
and more...


SIGN UP
Redemption
Recidivism
Departure

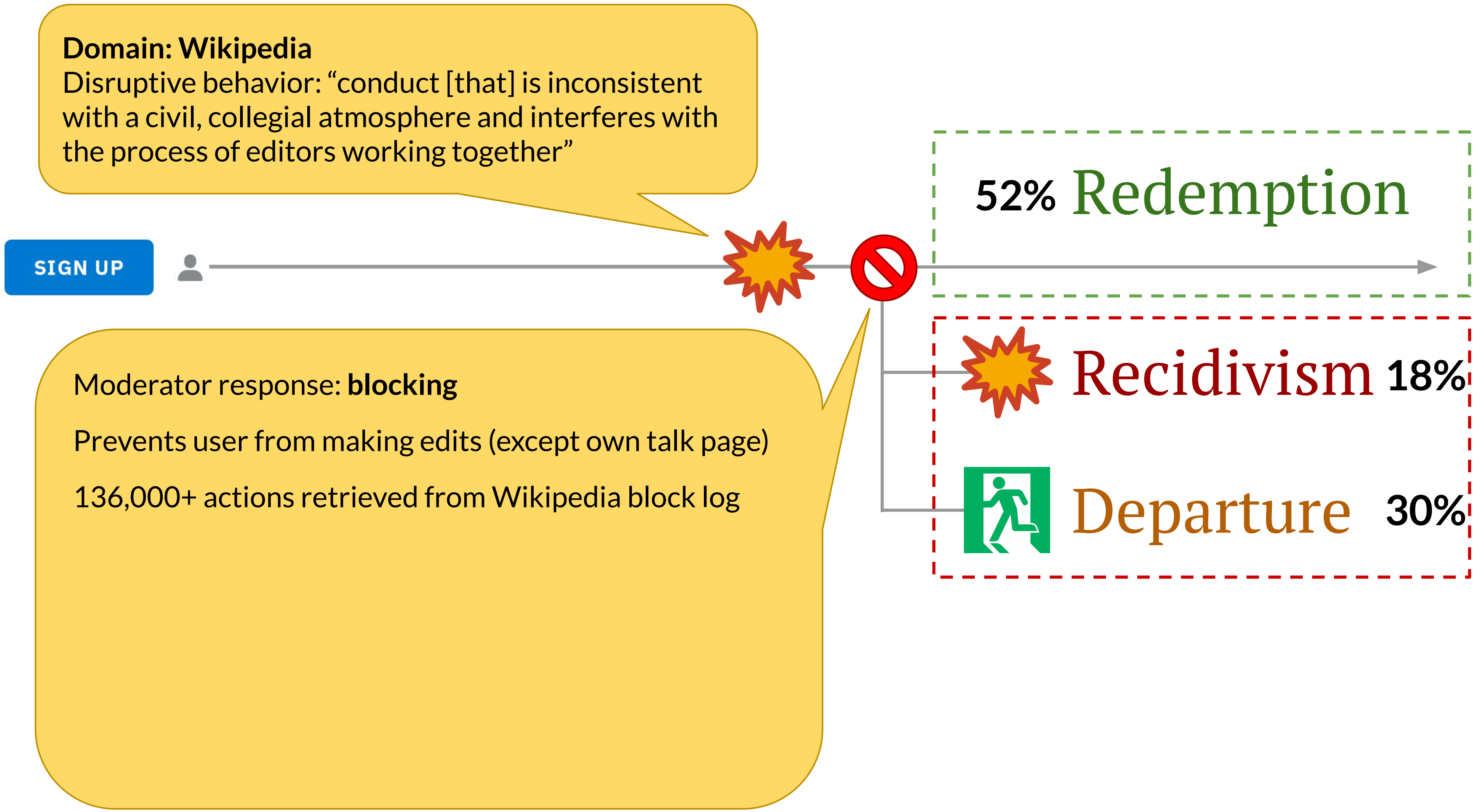

Domain: Wikipedia
Disruptive behavior: "conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together"
SIGN UP
Redemption
Recidivism
Departure

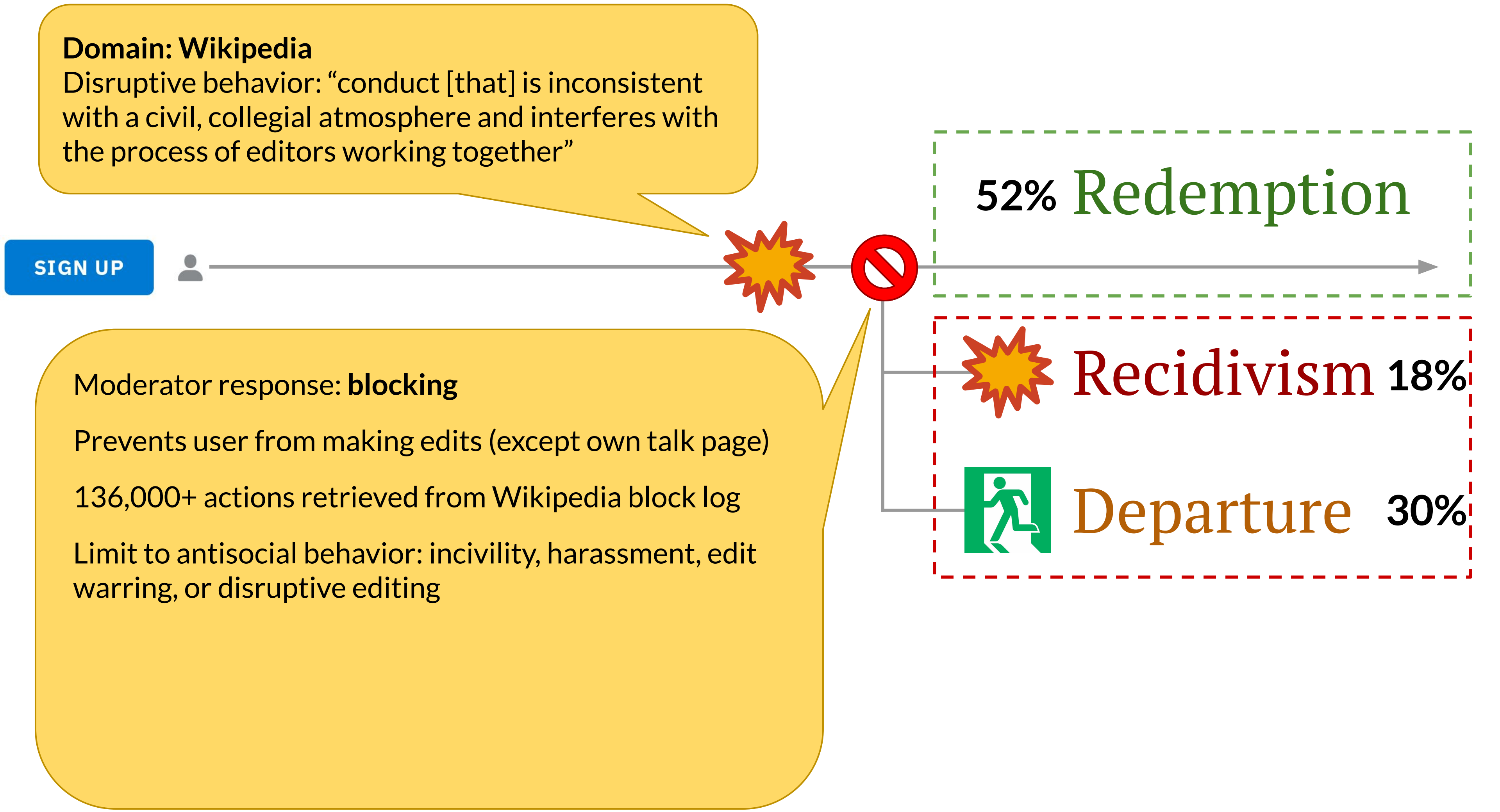


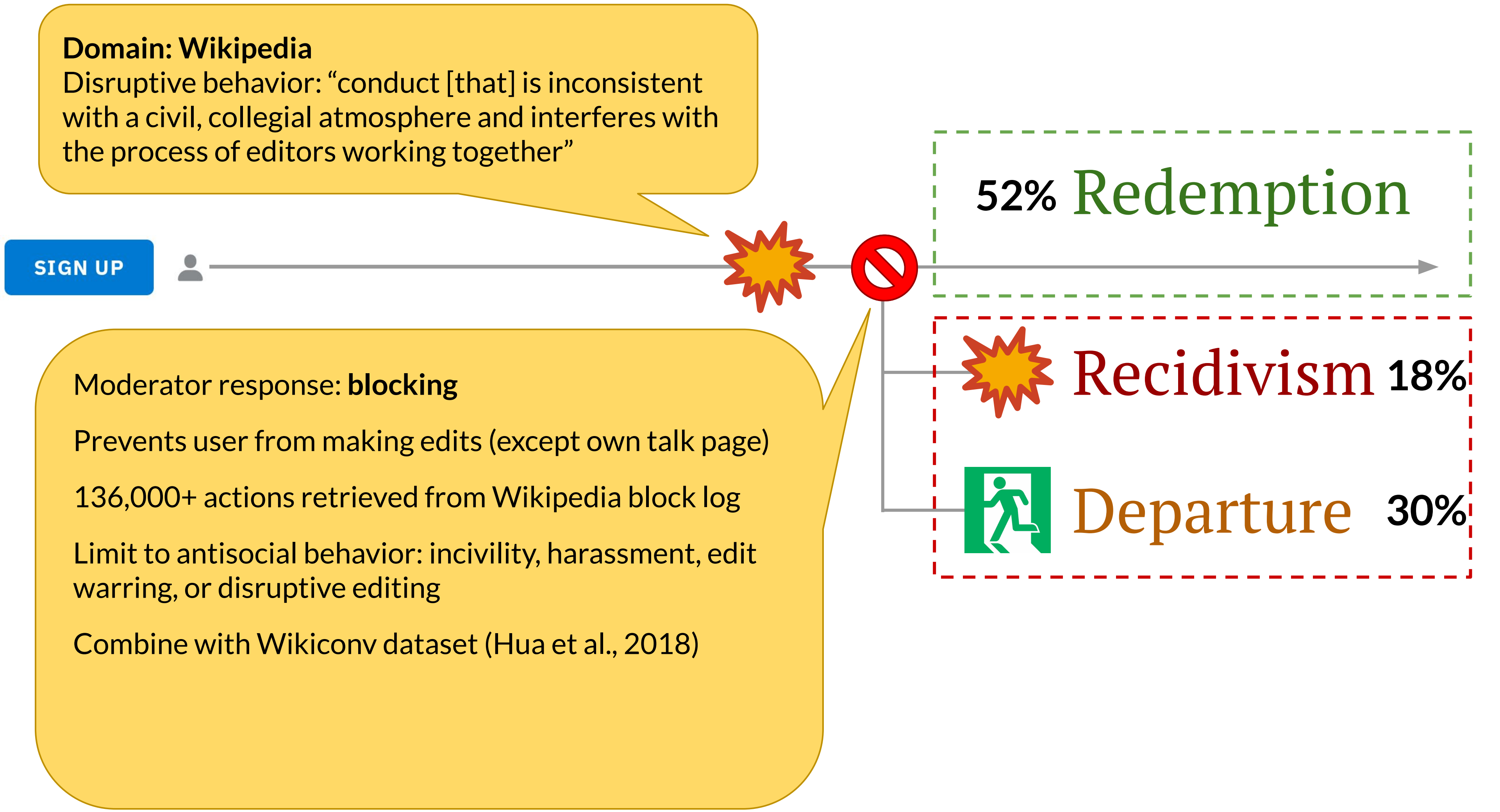
Domain: Wikipedia
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
SIGN UP
Redemption
Recidivism
Departure
Moderator response: blocking
Prevents user from making edits (except own talk page)



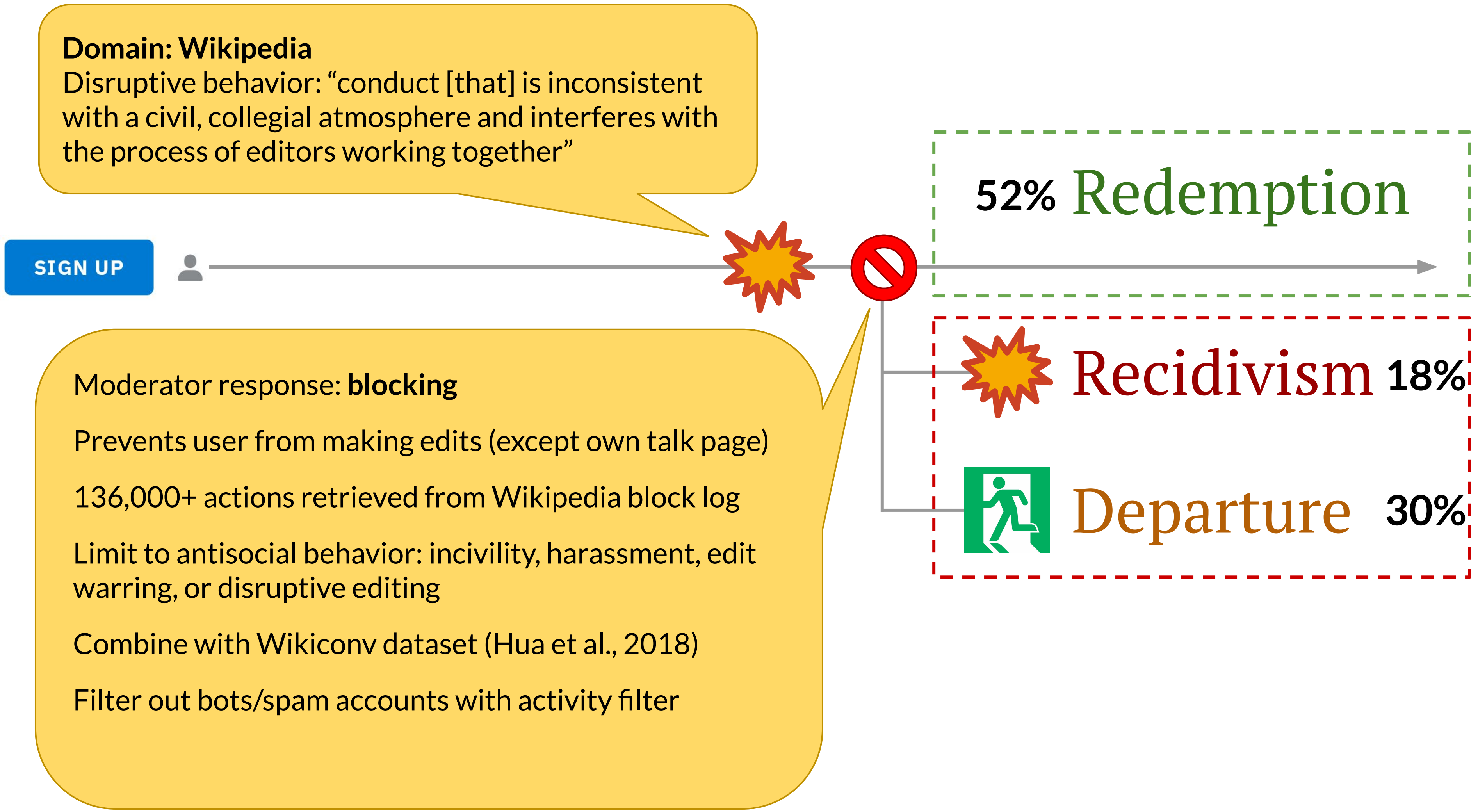
Domain: Wikipedia
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
SIGN UP
52% Redemption
Recidivism 18%
Departure 30%
Moderator response: blocking
Prevents user from making edits (except own talk page)

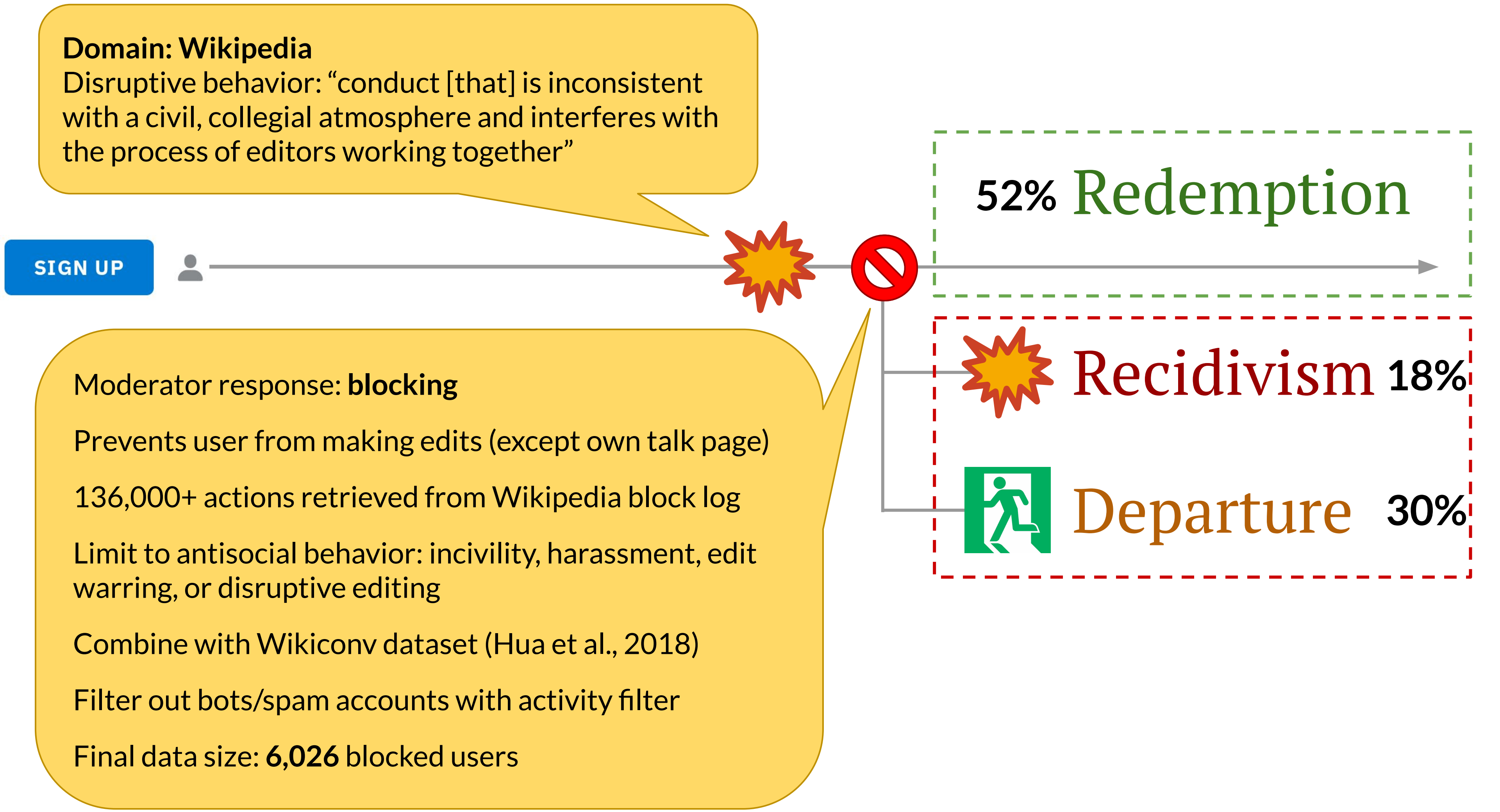

Domain: Wikipedia
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
SIGN UP
52% Redemption
Recidivism 18%
Departure 30%
Moderator response: **blocking**
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log


Domain: Wikipedia
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
SIGN UP
52% Redemption
Recidivism 18%
Departure 30%
Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing



Domain: Wikipedia
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
SIGN UP
52% Redemption
Recidivism 18%
Departure 30%
Moderator response: **blocking**
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing
Combine with Wikiconv dataset (Hua et al., 2018)


Domain: Wikipedia
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
SIGN UP
Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing
Combine with Wikiconv dataset (Hua et al., 2018)
Filter out bots/spam accounts with activity filter
52% Redemption
Recidivism 18%
Departure 30%



Domain: Wikipedia
Disruptive behavior: “conduct [that] is inconsistent with a civil, collegial atmosphere and interferes with the process of editors working together”
SIGN UP
52% Redemption
Recidivism 18%
Departure 30%
Moderator response: blocking
Prevents user from making edits (except own talk page)
136,000+ actions retrieved from Wikipedia block log
Limit to antisocial behavior: incivility, harassment, edit warring, or disruptive editing
Combine with Wikiconv dataset (Hua et al., 2018)
Filter out bots/spam accounts with activity filter
Final data size: 6,026 blocked users

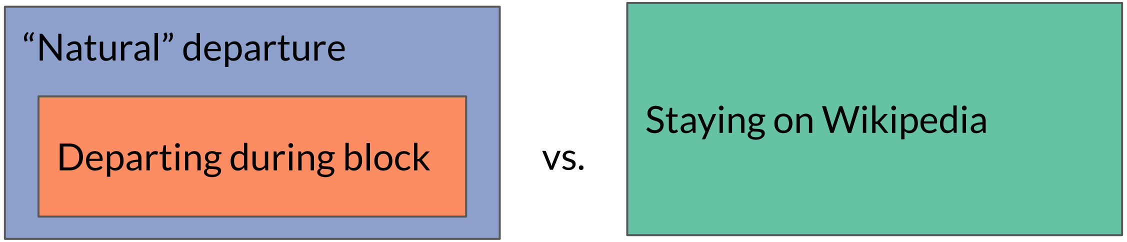
Making meaningful comparisons

Departing during block

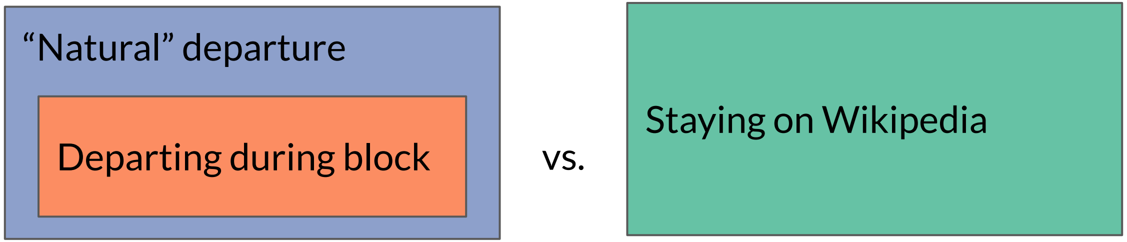
vs.

Staying on Wikipedia

Making meaningful comparisons



Making meaningful comparisons



Matching lets us make nontrivial comparisons between departure and redemption...

Making meaningful comparisons

“Natural” departure

Departing during block

vs.

Staying on Wikipedia

Matching lets us make nontrivial comparisons between departure and redemption...

D
(departs during block)

Making meaningful comparisons

“Natural” departure

Departing during block

vs.

Staying on Wikipedia

Matching lets us make nontrivial comparisons between departure and redemption...

Experimental Pair

D
(departs during block)

Reformed user (blocked around same time as **D**)

Making meaningful comparisons

“Natural” departure

Departing during block

vs.

Staying on Wikipedia

Matching lets us make nontrivial comparisons between departure and redemption...

...and a **second level** of matching ensures those comparisons are specific to block departures

Experimental Pair

D
(departs during block)

Reformed user (blocked around same time as **D**)

Making meaningful comparisons

“Natural” departure

Departing during block

vs.

Staying on Wikipedia

Matching lets us make nontrivial comparisons between departure and redemption...

...and a **second level** of matching ensures those comparisons are specific to block departures

Experimental Pair

D
(departs during block)

Reformed user (blocked around same time as **D**)

Control Pair

N
(never-blocked user, departs around same time as **D**)

Making meaningful comparisons

“Natural” departure

Departing during block

vs.

Staying on Wikipedia

Matching lets us make nontrivial comparisons between departure and redemption...

...and a **second level** of matching ensures those comparisons are specific to block departures

Experimental Pair

D (departs during block)

Reformed user (blocked around same time as **D**)

Control Pair

N (never-blocked user, departs around same time as **D**)

Clean user (active at/beyond time of **N**’s departure

Making meaningful comparisons

"Natural" departure

Departing during block

vs.

Staying on Wikipedia

Matching lets us make nontrivial comparisons between departure and redemption...

...and a **second level** of matching ensures those comparisons are specific to block departures

Experimental Pair

D (departs during block)

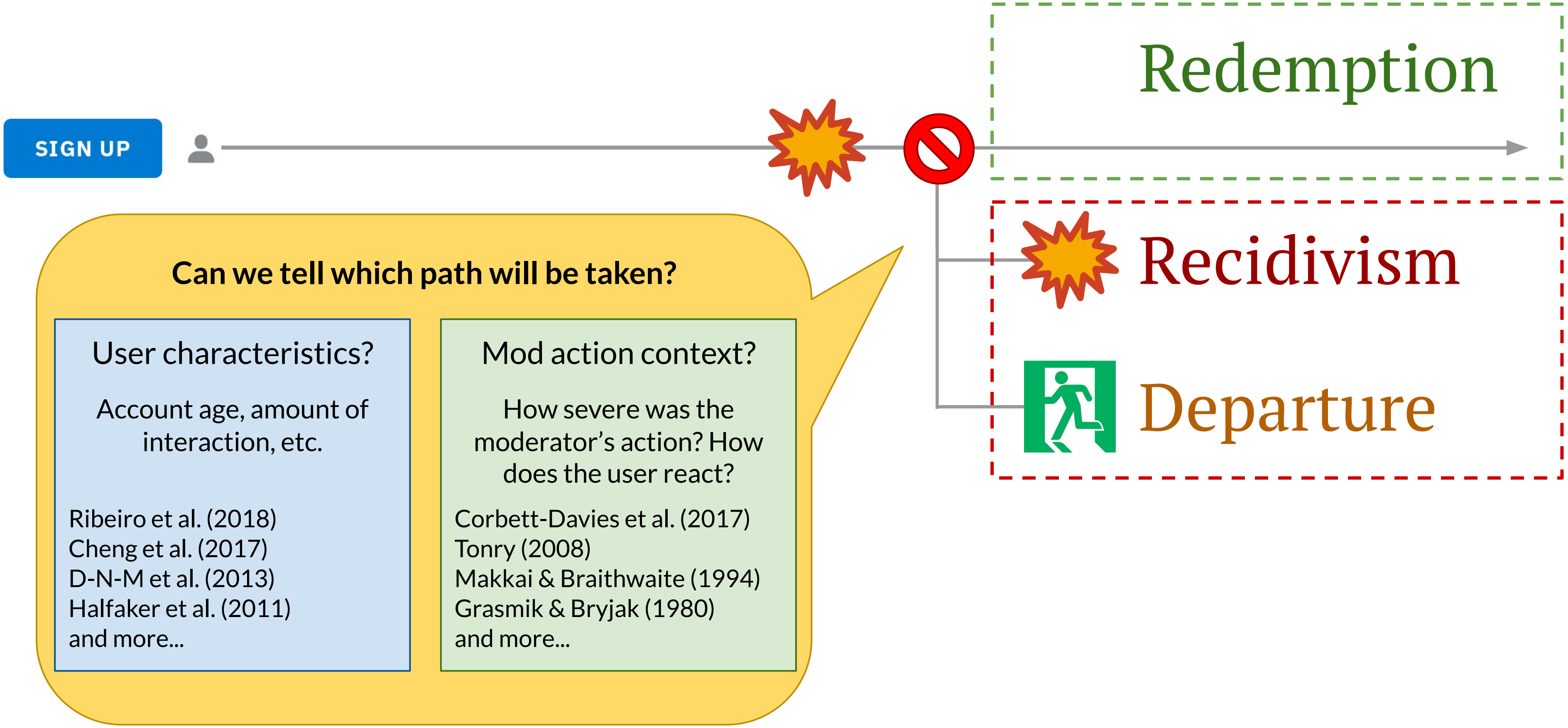
Reformed user (blocked around same time as **D**)

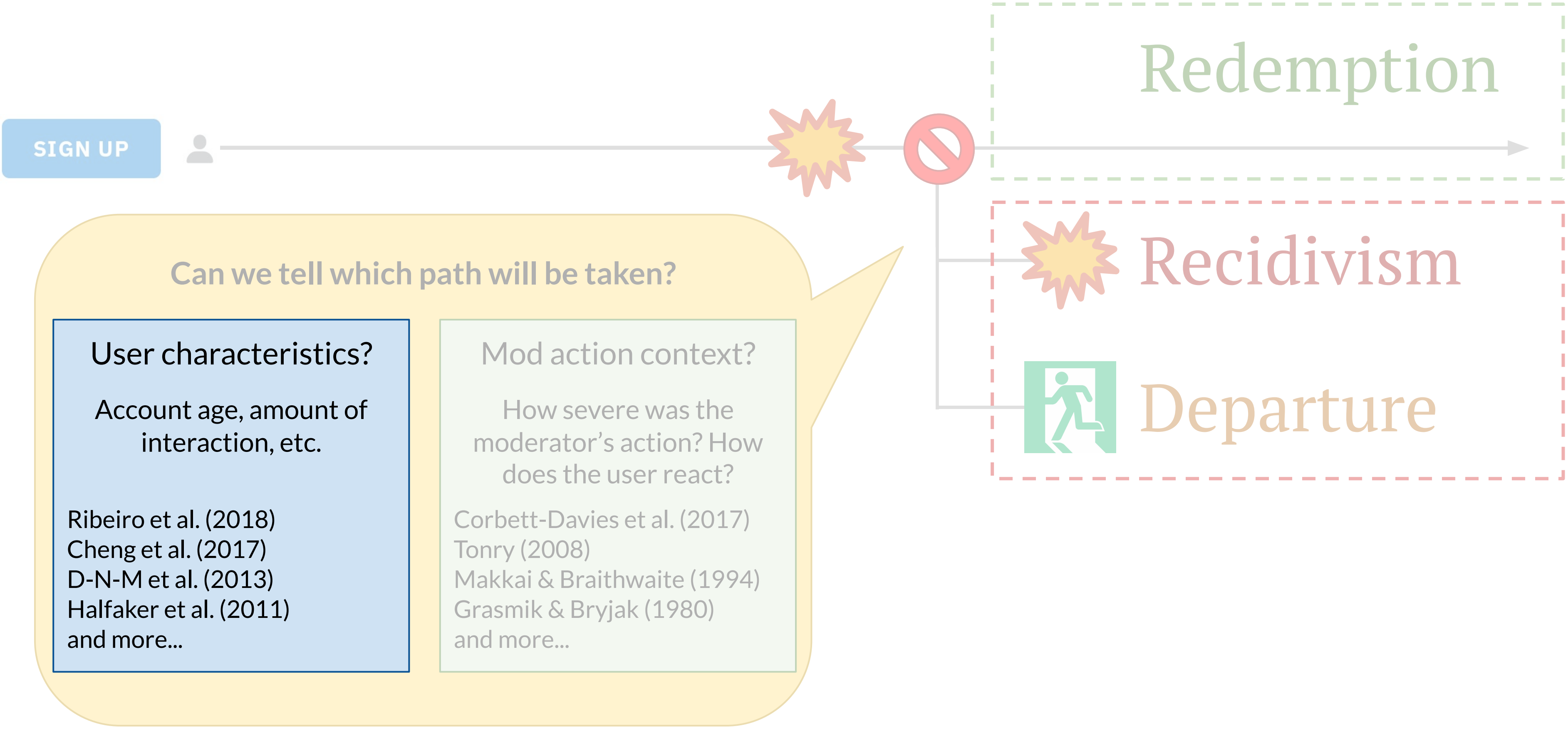
Control Pair

N (never-blocked user, departs around same time as **D**)

Clean user (active at/beyond time of **N**'s departure

(analogous process for recidivist vs reformed users)


SIGN UP
Redemption
Recidivism
Departure
Can we tell which path will be taken?
User characteristics?
Account age, amount of interaction, etc.
Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...
Mod action context?
How severe was the moderator's action? How does the user react?
Corbett-Davies et al. (2017)
Tonry (2008)
Makkai & Braithwaite (1994)
Grasmik & Bryjak (1980)
and more...


SIGN UP
Redemption
Recidivism
Departure
Can we tell which path will be taken?
User characteristics?
Account age, amount of interaction, etc.
Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...
Mod action context?
How severe was the moderator's action? How does the user react?
Corbett-Davies et al. (2017)
Tonry (2008)
Makkai & Braithwaite (1994)
Grasmik & Bryjak (1980)
and more...

What characteristics matter?

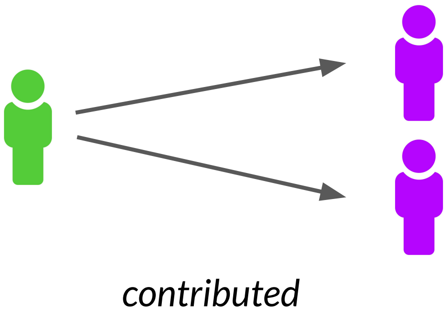
Prior work: norm violations correlate with level of involvement in community

Simple measure: number of talk page comments (“activity level”)

What characteristics matter?

Prior work: norm violations correlate with level of involvement in community

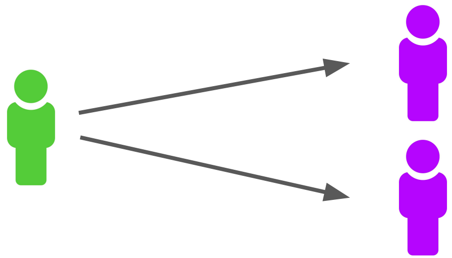
Simple measure: number of talk page comments (“activity level”)



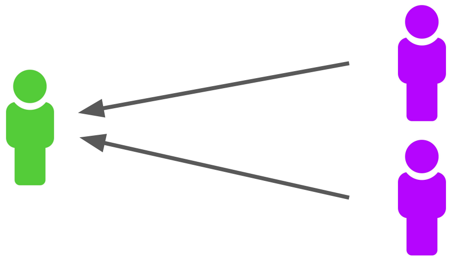
What characteristics matter?

Prior work: norm violations correlate with level of involvement in community

Simple measure: number of talk page comments (“activity level”)



contributed

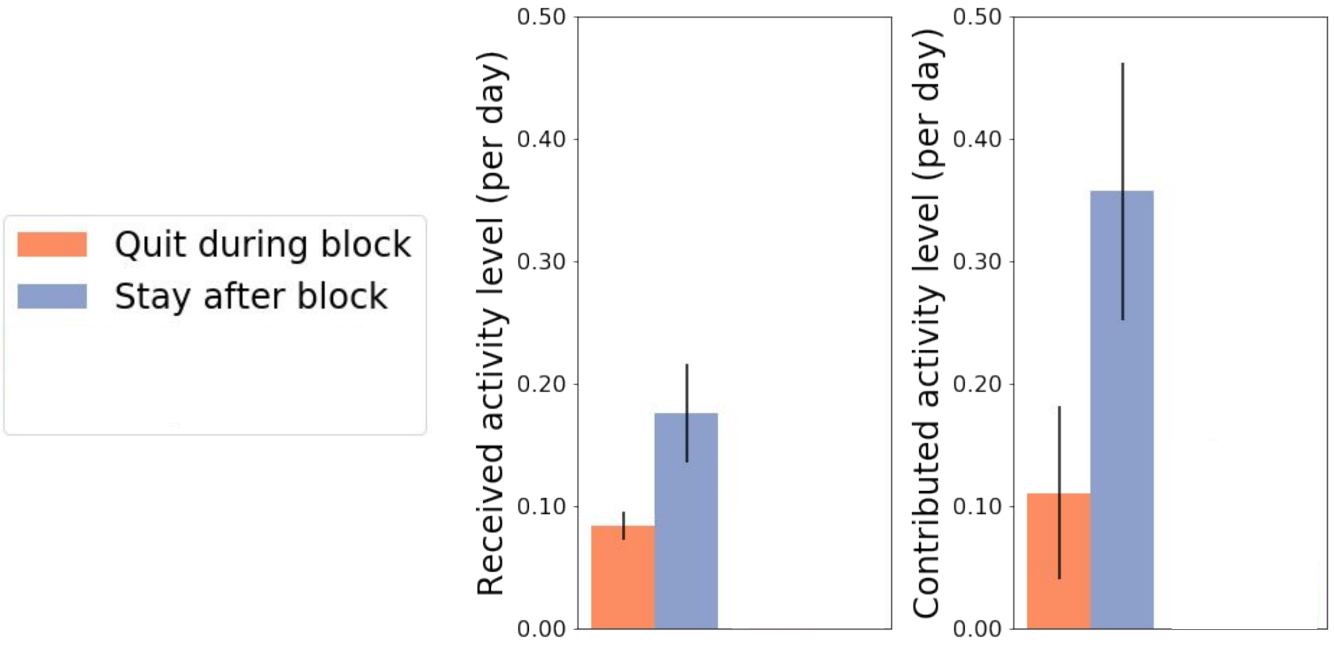


received

Activity level vs departure

Users who depart tend to have lower activity level

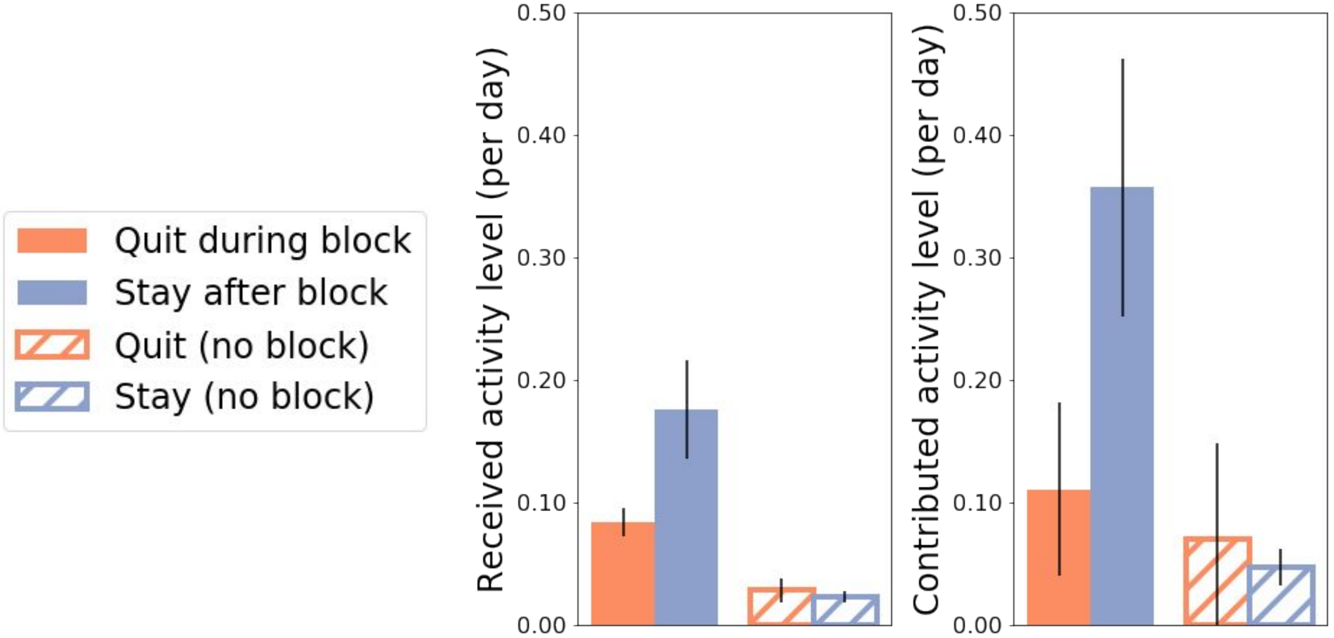
- Intuitive interpretation: less involvement → less reason to stay



Activity level vs departure

Users who depart tend to have lower activity level

- Intuitive interpretation: less involvement → less reason to stay



Beyond activity level

Activity level measures *amount* of involvement, but not *nature* of involvement

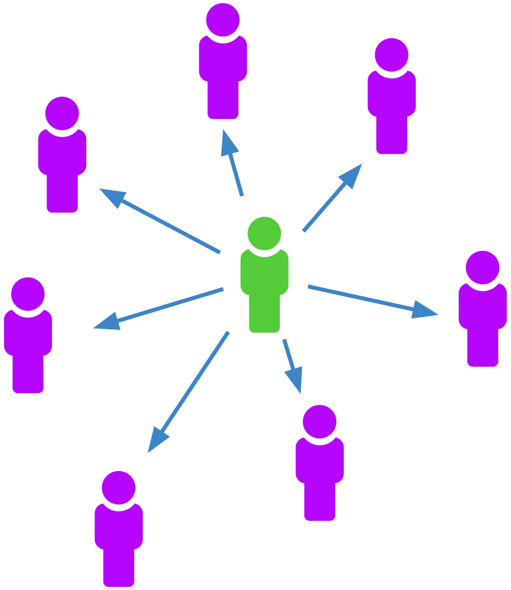
For the latter, need *activity spread*

Activity Spread

Example User has written 100 comments

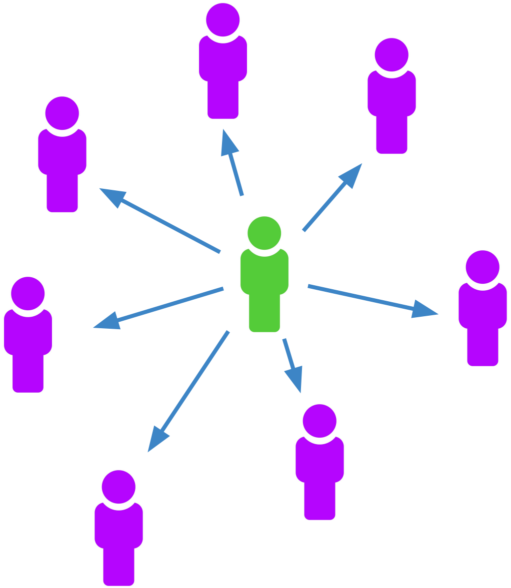
Activity Spread

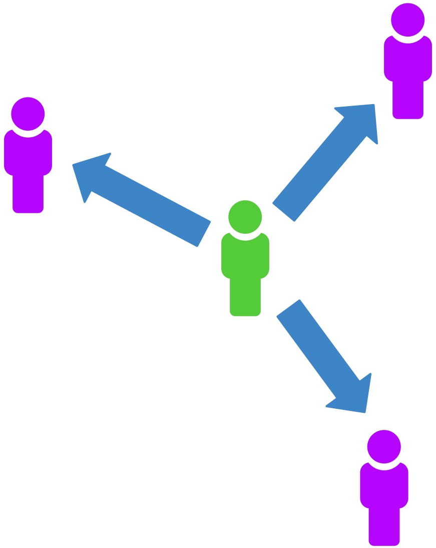
Example User has written 100 comments



Activity Spread

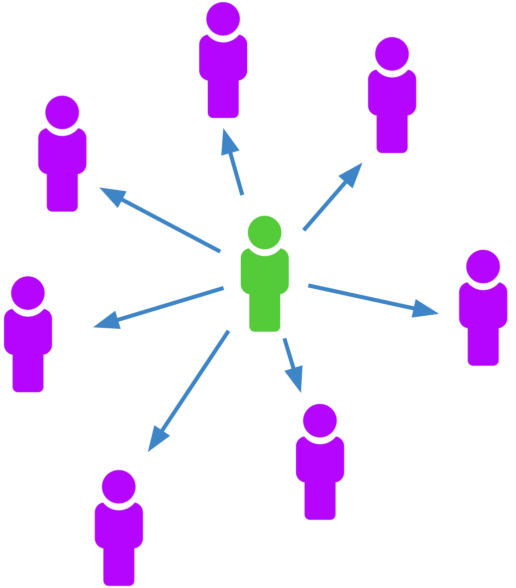
Example User has written 100 comments

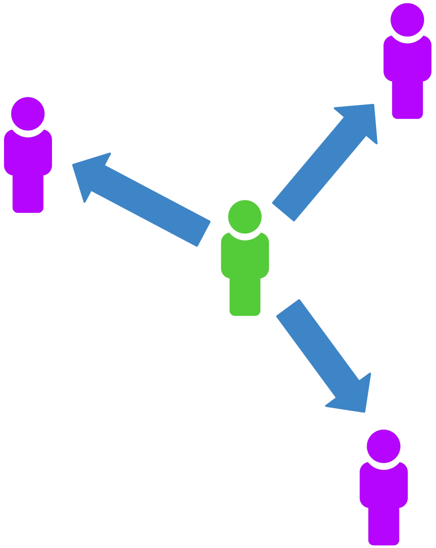




Activity Spread

Example User has written 100 comments

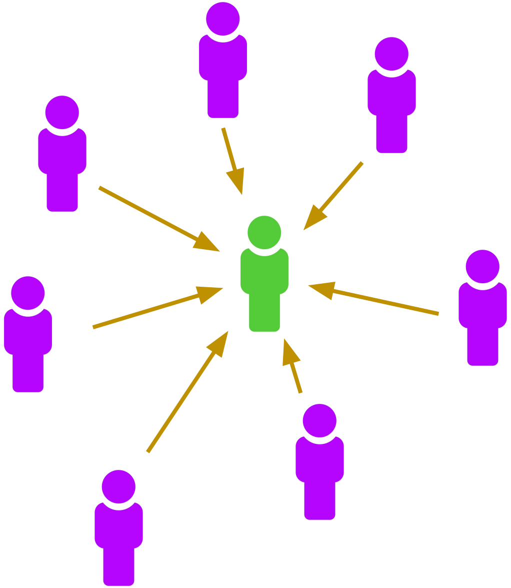


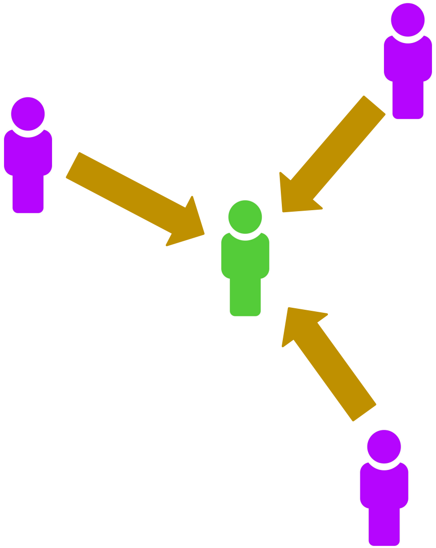


$$\text{contributed activity spread} = \frac{\#\ \text{comments written}}{\#\ \text{other user talk pages written to}}$$

Activity Spread

Example User has received 100 comments



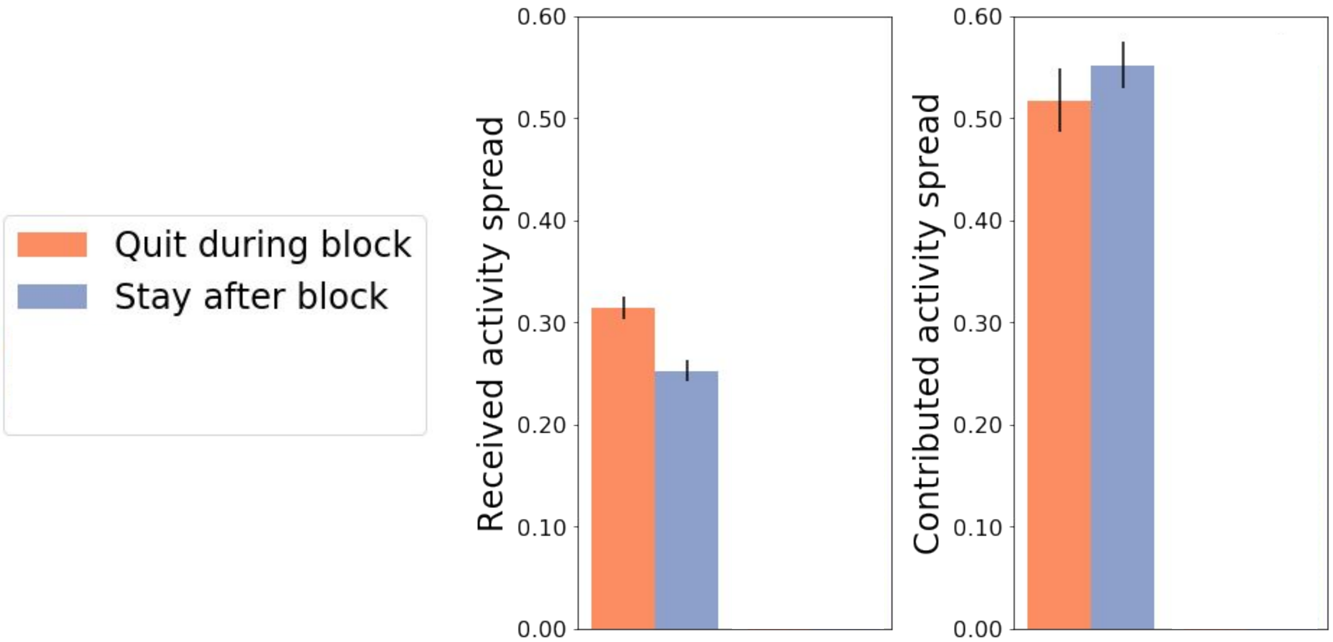


$$\text{received activity spread} = \frac{\text{\# comments received}}{\text{\# unique comment authors}}$$

Activity spread vs departure

Users who depart tend to have higher received activity spread

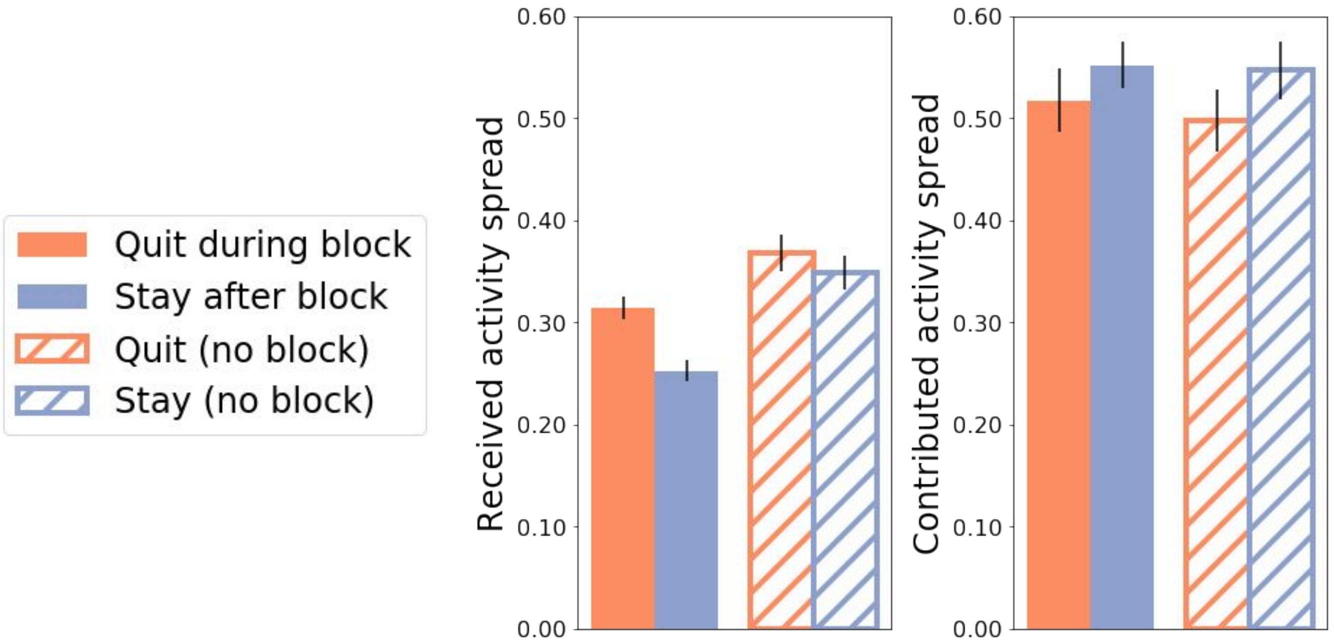
- Possible intuitive interpretation: less tightly integrated into a social circle



Activity spread vs departure

Users who depart tend to have higher received activity spread

- Possible intuitive interpretation: less tightly integrated into a social circle



Engagement features: predicting trajectories

Can the engagement measures be used to predict a blocked user's future trajectory?

Methodology: use engagement measures as features to SVM

- Separate models for predicting departure and recidivism

Baselines: block reason, block duration

Also consider how long user has been active ("community age")

- Basic measure of engagement

Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data
*s indicate significant ($p < 0.05$) improvement over best baseline in column

	Departure	Recidivism
Baseline: block reason		
Baseline: block duration		
Community age		
Engagement features		
Engagement + Age		

Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data
*s indicate significant ($p < 0.05$) improvement over best baseline in column

	Departure	Recidivism
Baseline: block reason	59.0	
Baseline: block duration	56.7	
Community age		
Engagement features		
Engagement + Age		

Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data
*s indicate significant ($p < 0.05$) improvement over best baseline in column

	Departure	Recidivism
Baseline: block reason	59.0	
Baseline: block duration	56.7	
Community age	58.6	
Engagement features	61.4*	
Engagement + Age		

Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data
*s indicate significant ($p < 0.05$) improvement over best baseline in column

	Departure	Recidivism
Baseline: block reason	59.0	
Baseline: block duration	56.7	
Community age	58.6	
Engagement features	61.4*	
Engagement + Age	66.2*	

Engagement features: predicting trajectories

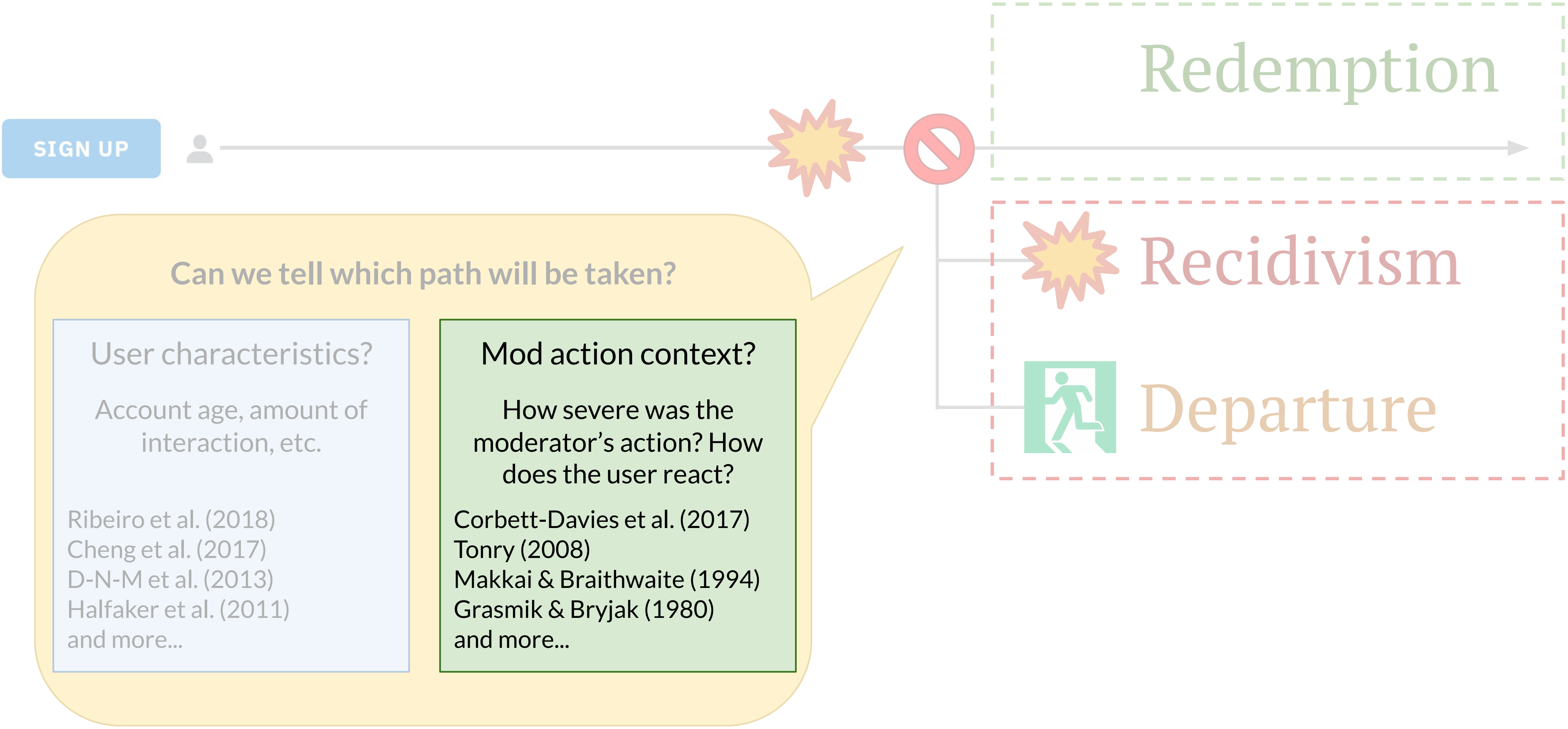
All accuracies are computed via leave-one-out CV on balanced (paired) data
*s indicate significant ($p < 0.05$) improvement over best baseline in column

	Departure	Recidivism
Baseline: block reason	59.0	51.9
Baseline: block duration	56.7	43.8
Community age	58.6	
Engagement features	61.4*	
Engagement + Age	66.2*	

Engagement features: predicting trajectories

All accuracies are computed via leave-one-out CV on balanced (paired) data
*s indicate significant ($p < 0.05$) improvement over best baseline in column

	Departure	Recidivism
Baseline: block reason	59.0	51.9
Baseline: block duration	56.7	43.8
Community age	58.6	56.3*
Engagement features	61.4*	59.1*
Engagement + Age	66.2*	58.8*


SIGN UP
Redemption
Recidivism
Departure
Can we tell which path will be taken?
User characteristics?
Account age, amount of interaction, etc.
Ribeiro et al. (2018)
Cheng et al. (2017)
D-N-M et al. (2013)
Halfaker et al. (2011)
and more...
Mod action context?
How severe was the moderator's action? How does the user react?
Corbett-Davies et al. (2017)
Tonry (2008)
Makkai & Braithwaite (1994)
Grasmik & Bryjak (1980)
and more...

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

Likelihood of repeat offense depends on...

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

Likelihood of repeat offense depends on...

...severity of the punishment

Grasmick and Bryjack (1980); Klepper and Nagin (1989)

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

Likelihood of repeat offense depends on...

...severity of the punishment

...offender's perception of punishment as fair

Grasmick and Bryjack (1980); Klepper and Nagin (1989)

Makkai and Braithwaite (1994); Paternoster et al. (1997); Williams (2005); Tonry (2008)

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

Likelihood of repeat offense depends on...

...severity of the punishment

...offender's perception of punishment as fair

Block context: measuring perceived fairness

2 angles: *blocked user's* perspective and *admin's* perspective

What can blocked users do to signal that they view block as (un)fair?

What can admins do to signal to blocked users that rules are fair?

Block context: measuring perceived fairness

2 angles: *blocked user's* perspective and *admin's* perspective

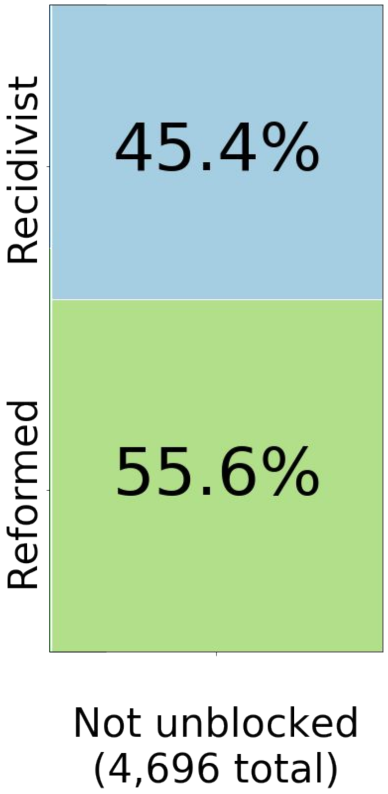
What can blocked users do to signal that they view block as (un)fair?

- Talk page comments

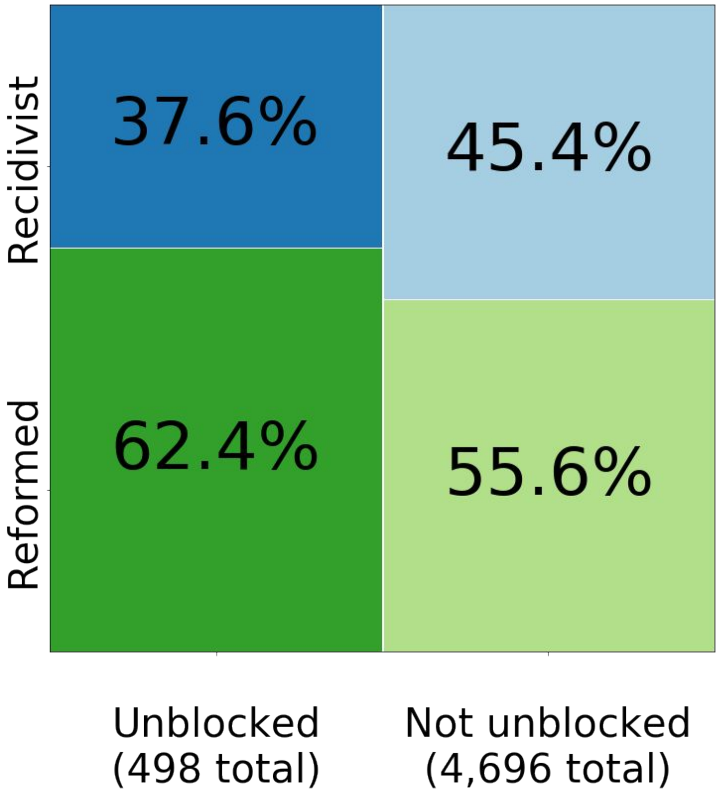
What can admins do to signal to blocked users that rules are fair?

- Unblocks

Admin's perspective: Unblocks



Admin's perspective: Unblocks



User’s perspective: Perceived fairness

3 linguistic indicators of user’s perception of fairness in comments:

Apologizing (suggests user acknowledges fairness of block)

- e.g., “I am deeply **sorry** for not understanding the whole situation, and ask for your forgiveness.”

Direct questioning (hostile; suggests user is fighting back)

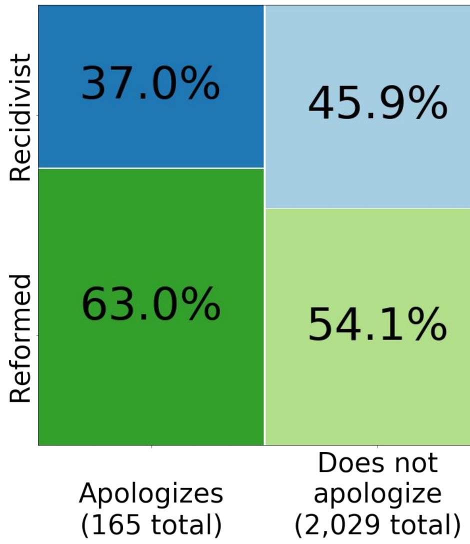
- e.g., “**So what** policy, precisely have I violated?”

Explicit mentions of “unfairness” and related phrases

- e.g., “i have alerted another administrator about your blatent [sic] and **unwarranted** abuse of power”

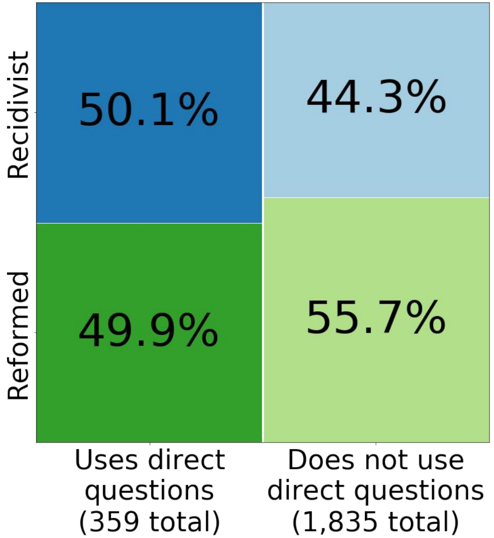
User's perspective: Perceived fairness

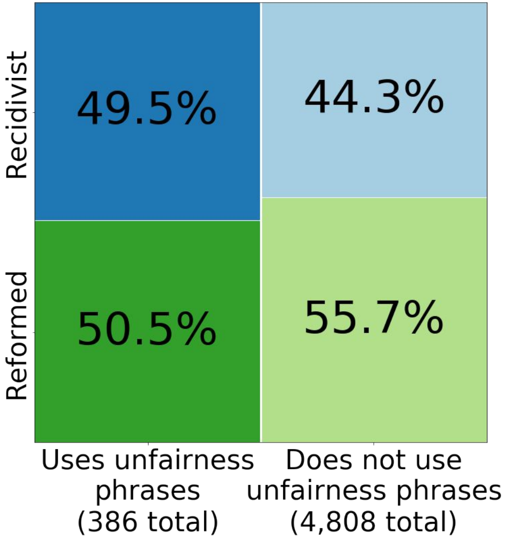
Likelihood of recidivism is lower for users who apologize



User's perspective: Perceived fairness

Likelihood of recidivism is higher for users who use unfairness phrases or direct questioning





Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

Likelihood of repeat offense depends on...

~~...severity of the punishment~~

...offender's perception of punishment as fair

Block context: motivation

How do properties of *first* block affect likelihood of *another* block?

Unique to recidivism (as opposed to getting blocked in general)

Loosely motivated by studies in offline law compliance

2 views of offline recidivism:

Concluding Thoughts

Limitations / Future work

What **not** to do: deploy this classifier in production setting

- Not trained for a realistic setting!
- Even with really “good” classifier, would require careful analysis of potential biases and other risks

Limitations / Future work

What **not** to do: deploy this classifier in production setting

- Not trained for a realistic setting!
- Even with really “good” classifier, would require careful analysis of potential biases and other risks

Drawbacks of setup

- “Unfairness lexicon” is incomplete, crude measure
- Departure is not binary!
- Lack of another block doesn’t necessarily mean lack of reoffense

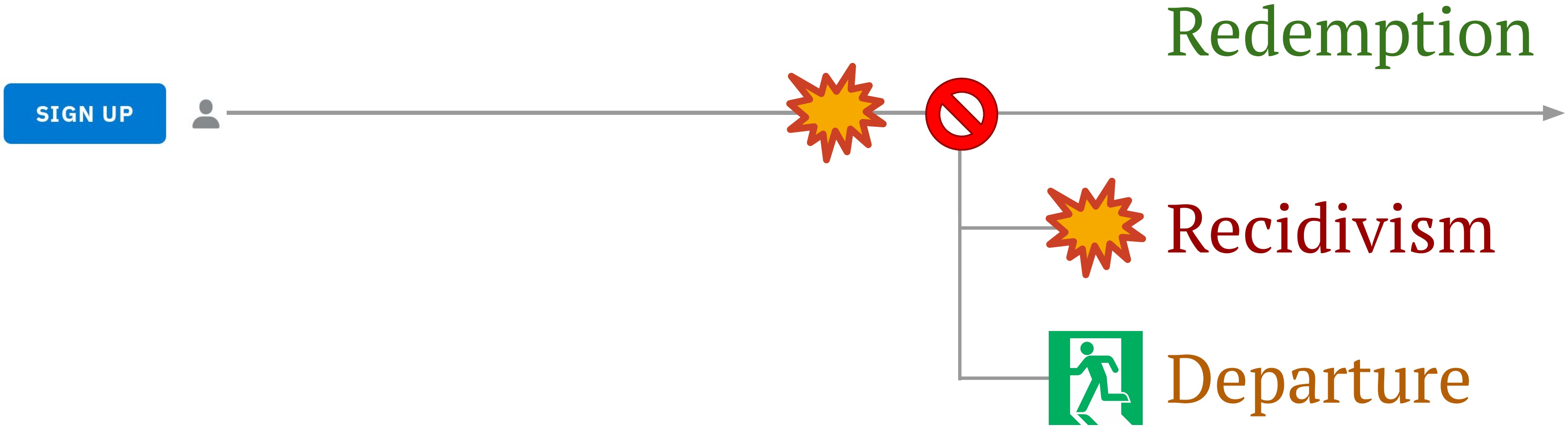
Limitations / Future work

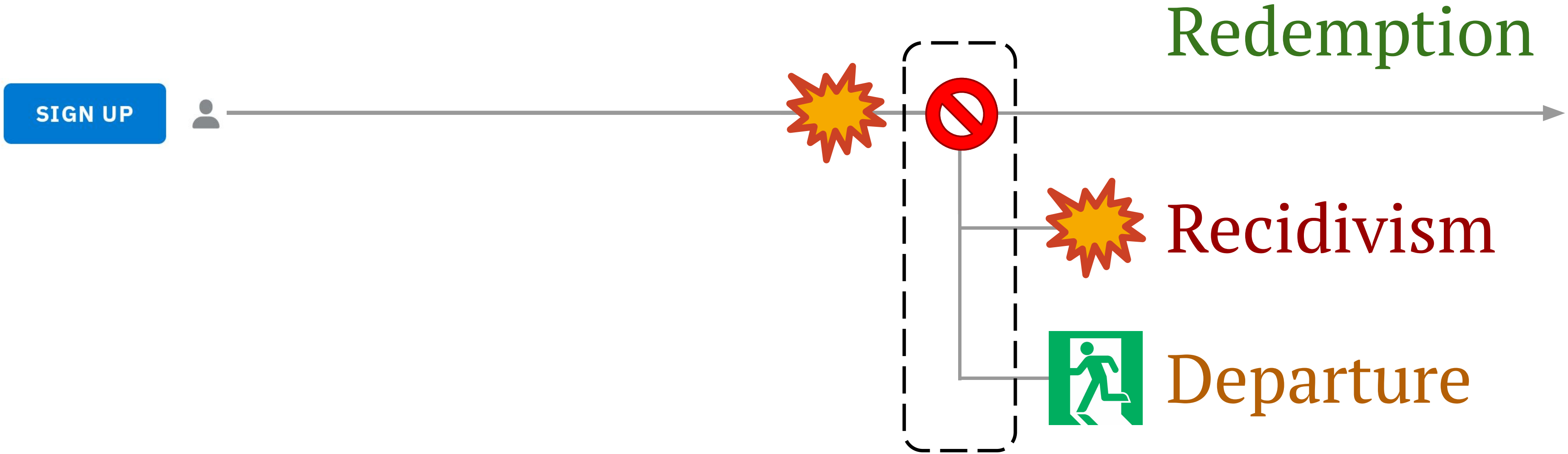
What **not** to do: deploy this classifier in production setting

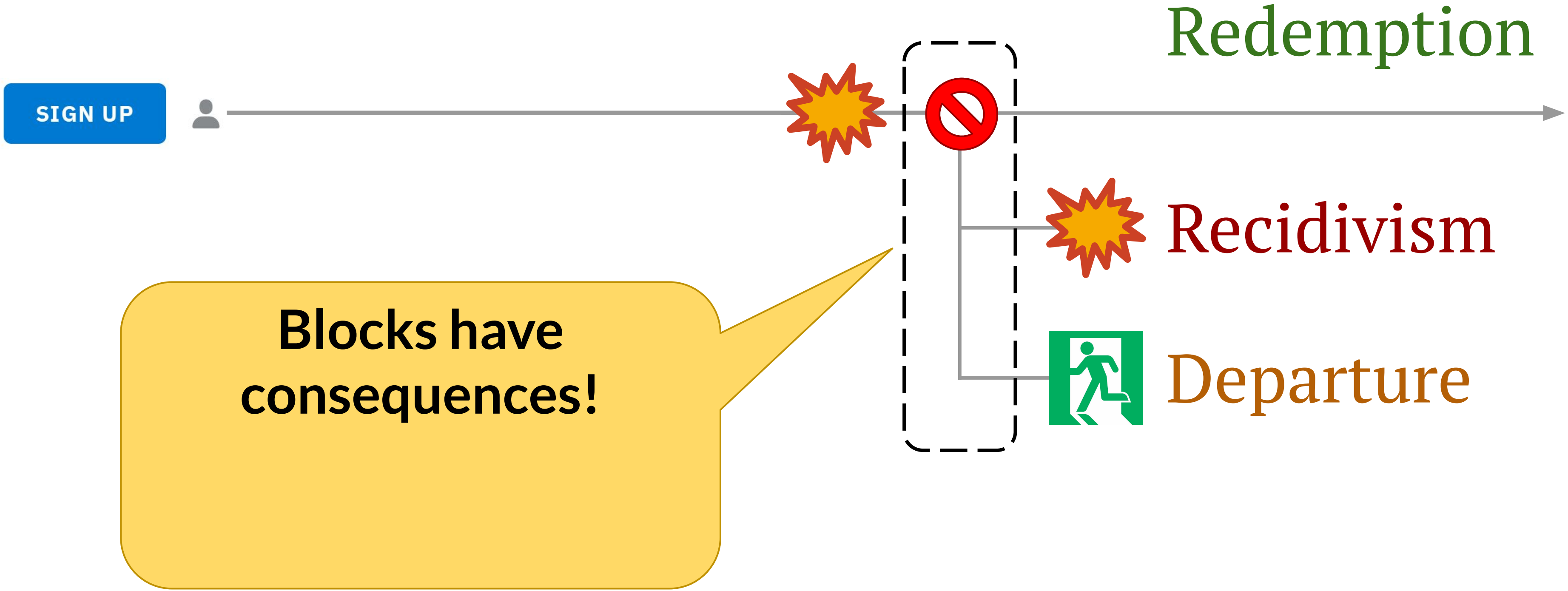
- Not trained for a realistic setting!
- Even with really “good” classifier, would require careful analysis of potential biases and other risks

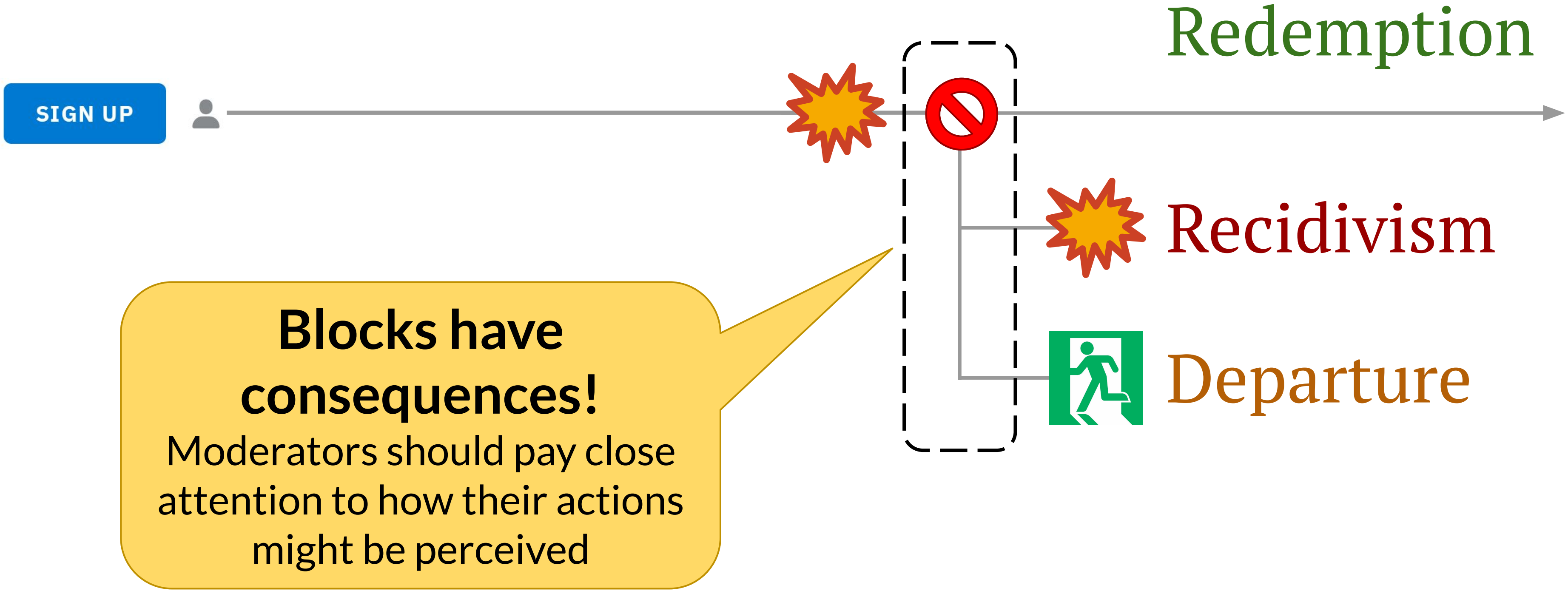
Drawbacks of setup

- “Unfairness lexicon” is incomplete, crude measure
- Departure is not binary!
- Lack of another block doesn’t necessarily mean lack of reoffense


SIGN UP
Redemption
Recidivism
Departure


SIGN UP
Redemption
Recidivism
Departure


SIGN UP
Redemption
Recidivism
Departure
Blocks have consequences!


SIGN UP
Redemption
Recidivism
Departure
Blocks have consequences!
Moderators should pay close attention to how their actions might be perceived

Questions?

Code and data available via ConvoKit (convokit.cornell.edu)