

# TOPIC MODELING IN THE ASEXUAL CENSUS

by Bauer McClave  
3/3/2016

# GOALS

- Describe topics of long form responses in survey data
- Find key words and phrases for each topic
- Create dictionary lookup application with Word2Vec

# USE CASE:

- Current survey report is static
- Data analysis done by volunteers with no reporting on long form answers so far
- Allows community members to investigate for themselves



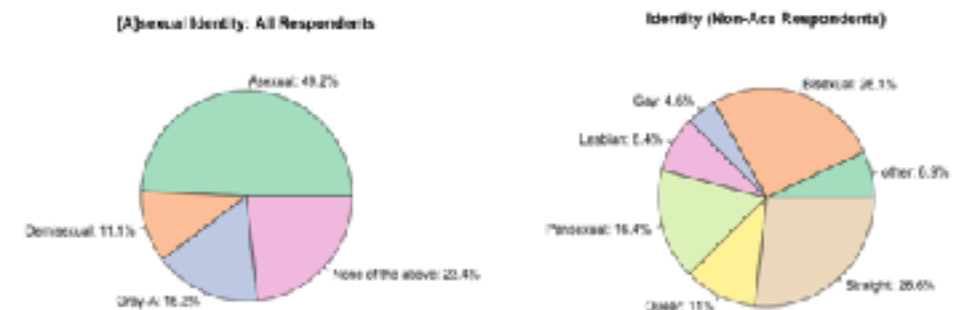
## 2014 AVEN Community Census: Preliminary Findings

### Section 1: General Survey Demographics

Unless otherwise specified, results in this section include all survey respondents - both aces and non-aces.

#### 1.1 Sexual Orientation

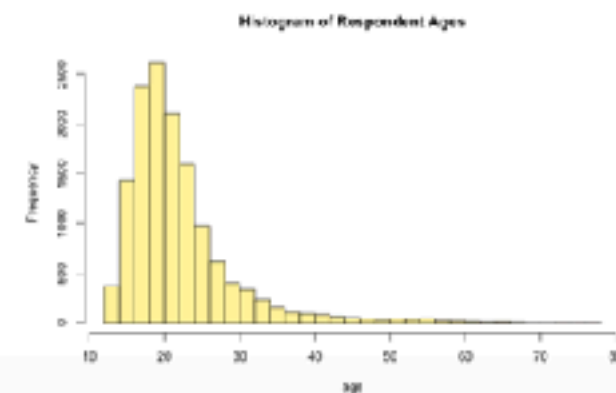
For the purposes of this survey, those who chose "Asexual", "Grey-Asexual", or "Demisexual" have been classed as "Aces", while those who chose "None of the Above" were classed as "non-aces". Those who chose "None of the Above" were given an additional set of sexual orientations and asked again to choose which they most identified with.



(Note: for breakdowns of romantic orientation, see the Asexual Identity section.)

#### 1.2 Age

Respondents ranged in age from 13-77, with a median age of 21 and mean age of 22.45. (The survey was limited to respondents over age 13 for legal reasons).



# AVEN COMMUNITY CENSUS

- Asexual Census - community of people who do not experience sexual attraction (with much nuance)
  - 9,850 with 6,311 respondents wrote long answer
    - Tell others? Their response?
    - Experience prejudice or discrimination
    - Support/services needed from asexual community
    - Support/services needed from other communities
    - Unanswered questions about the community
    - Other stories you would like to share

# METHODS

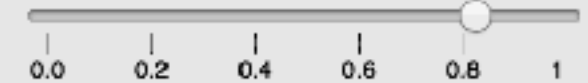
- Survey Creation team member (Oct. 2016)
- Pre-processing
- Topic Modeling:
  - Count Vectorizer
  - LDA
  - Word2vec & TSNE



# WHAT KIND OF SUPPORT / SERVICES DO YOU MOST NEED FROM ACE COMMUNITIES?

Selected Topic:

Slide to adjust relevance metric:(2)  
 $\lambda = 0.83$

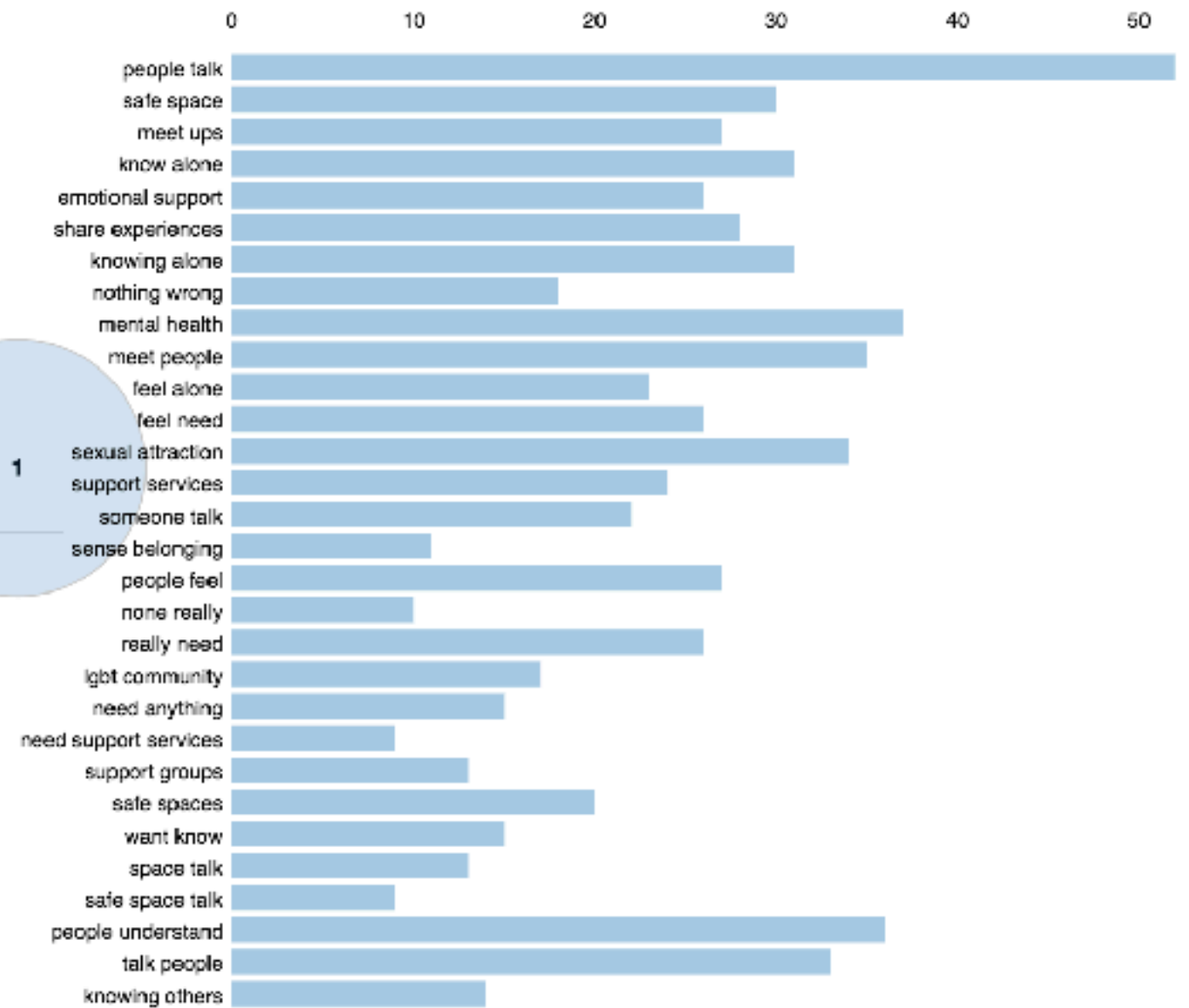


Intertopic Distance Map (via multidimensional scaling)



**Ace Community**

Top-30 Most Salient Terms<sup>1</sup>



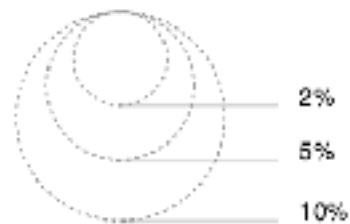
Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

Marginal topic distribution

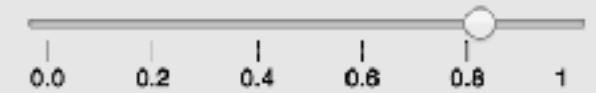




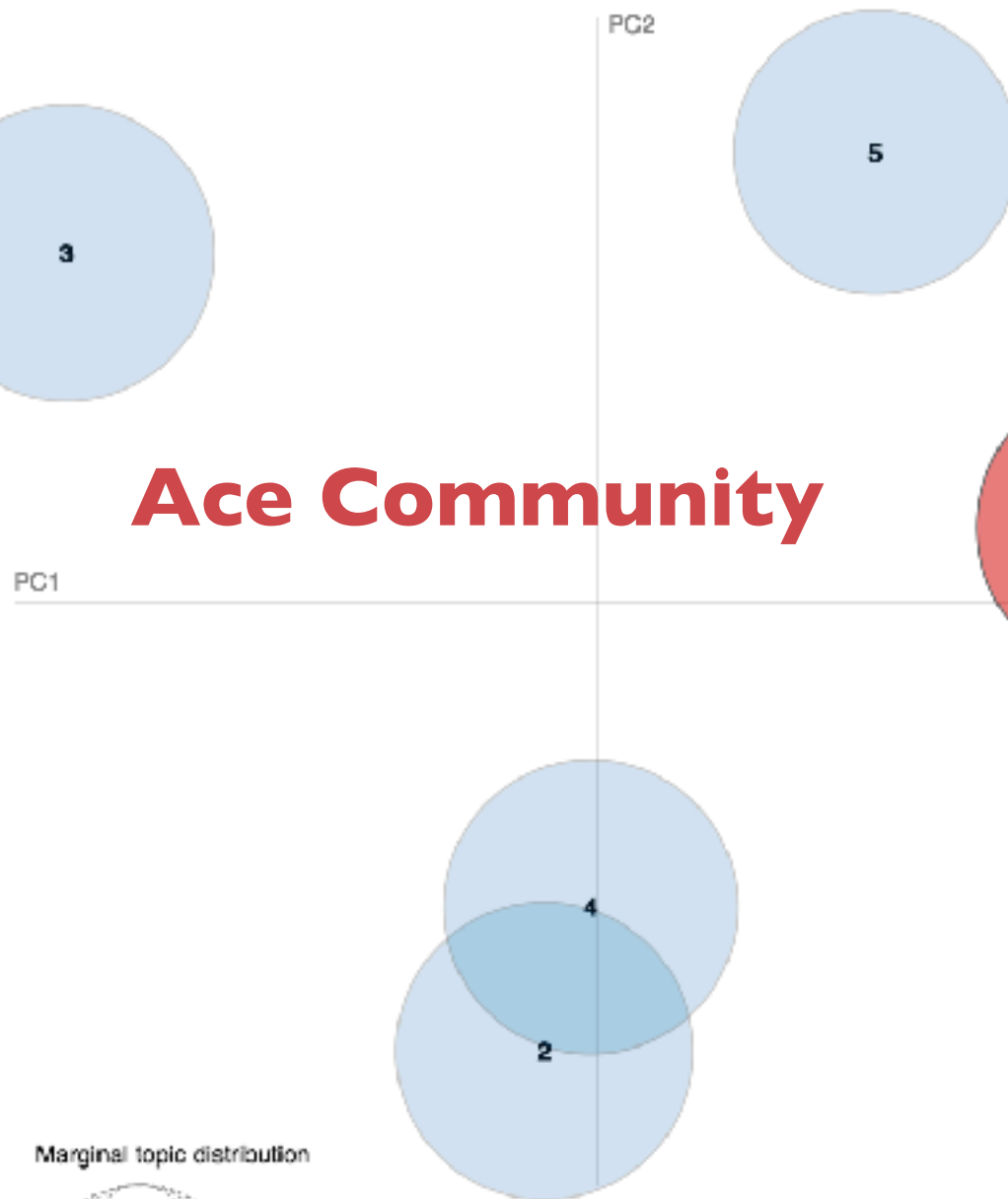
# WHAT KIND OF SUPPORT / SERVICES DO YOU MOST NEED FROM ACE COMMUNITIES?

Selected Topic:  Previous Topic Next Topic Clear Topic

Slide to adjust relevance metric: (2)  
 $\lambda = 0.83$



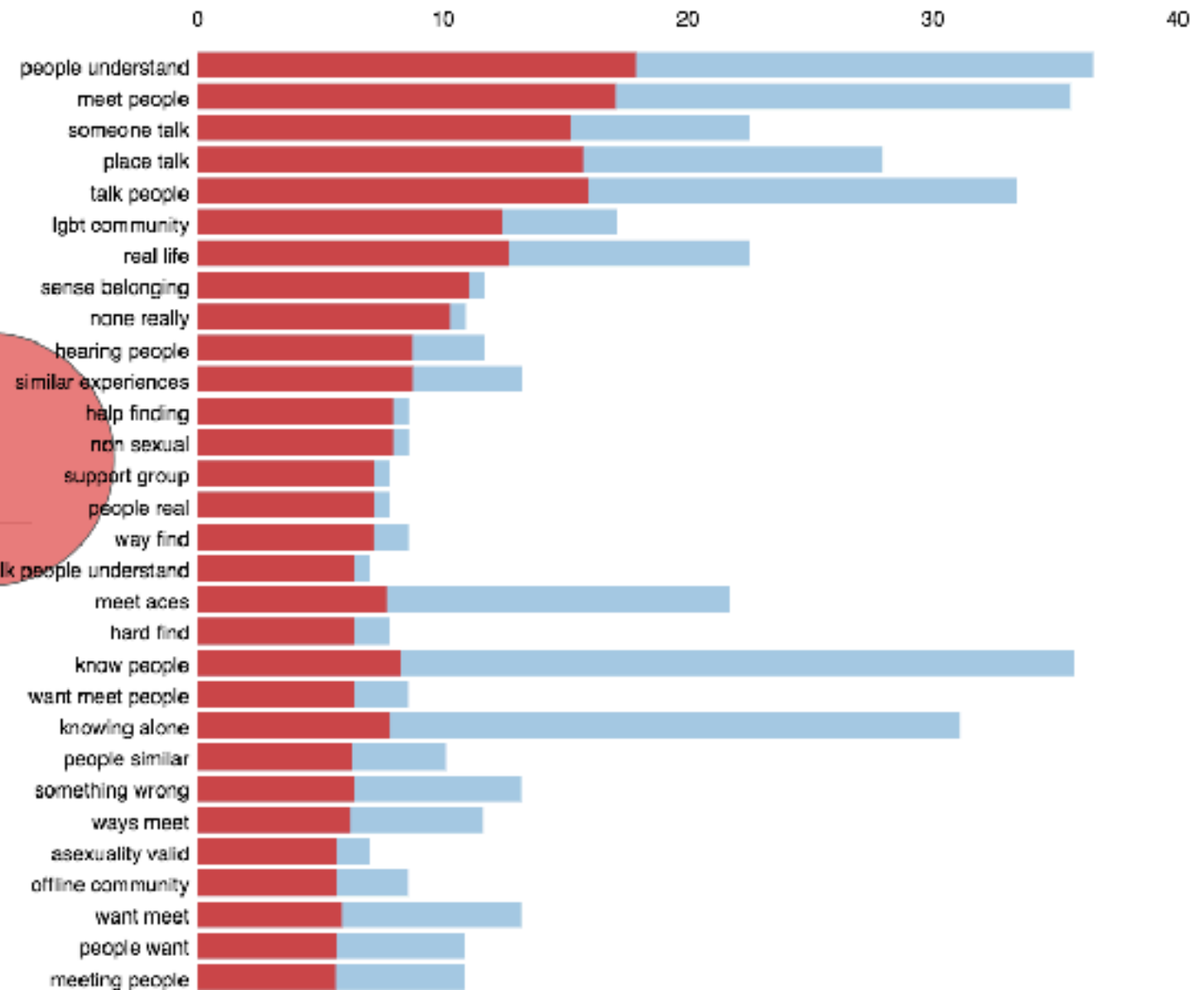
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 1 (20.9% of tokens)



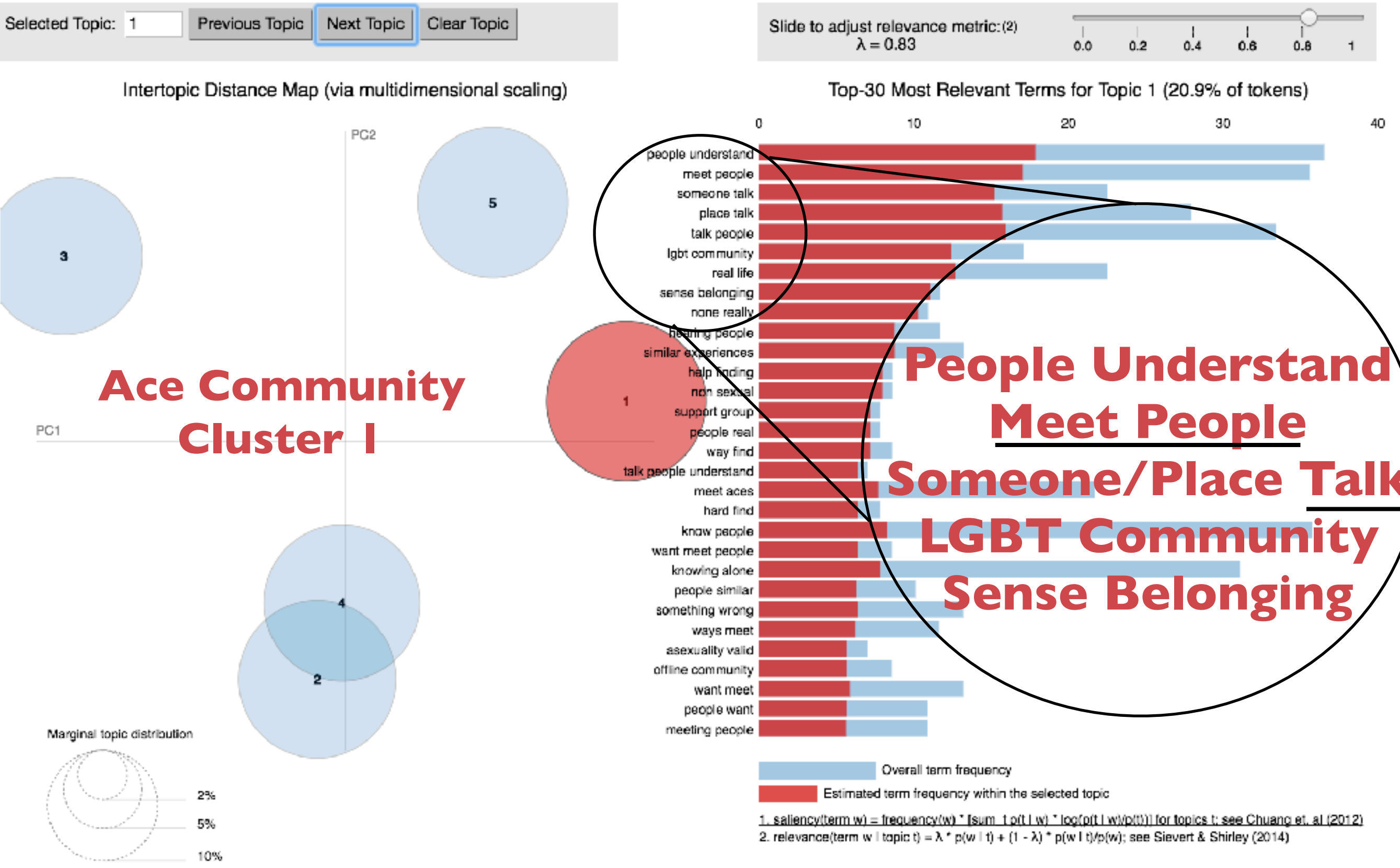
Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) \* [sum\_t p(t | w) \* log(p(t | w)/p(t))]; see Chuang et. al (2012)

2. relevance(term w | topic t) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

# WHAT KIND OF SUPPORT / SERVICES DO YOU MOST NEED FROM ACE COMMUNITIES?

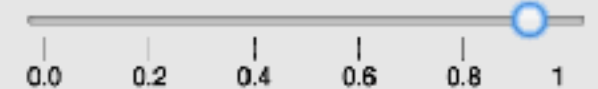




# WHAT KIND OF SUPPORT / SERVICES DO YOU MOST NEED FROM OTHER COMMUNITIES?

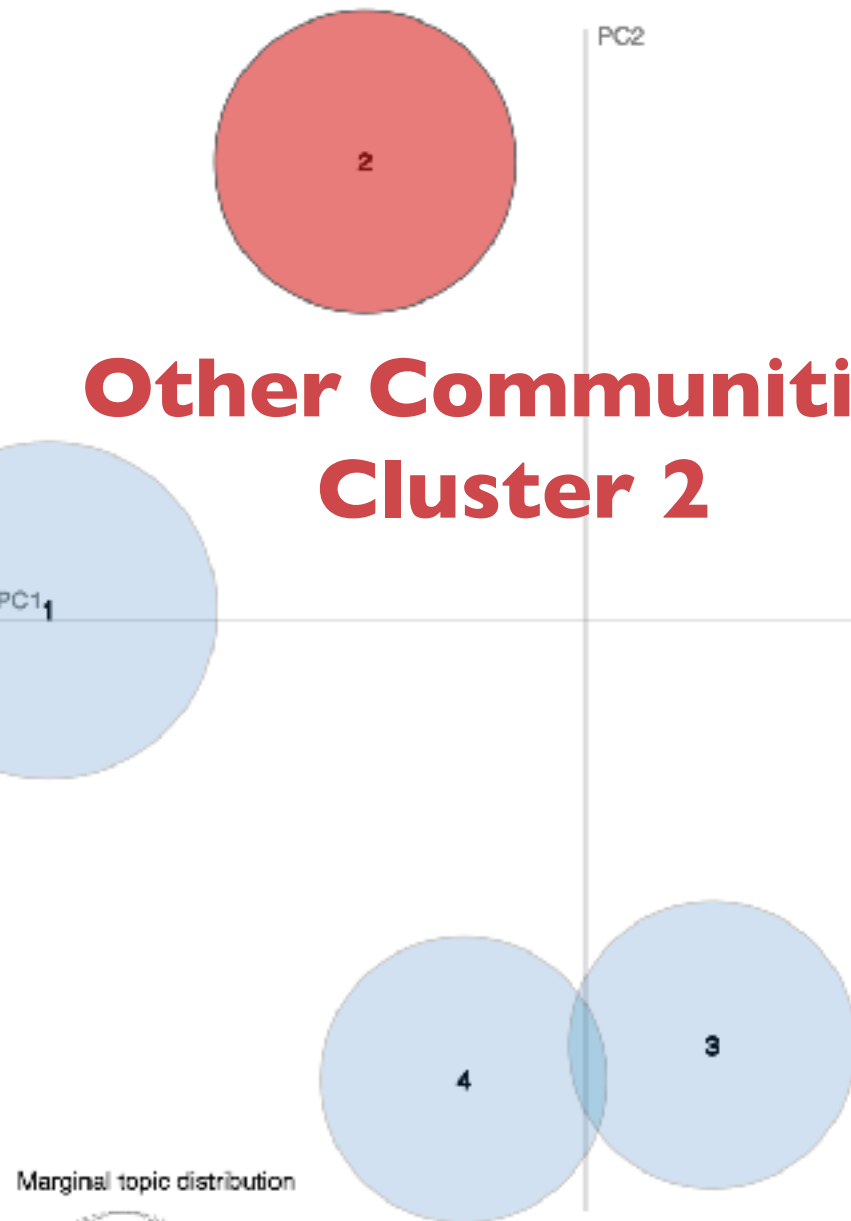
Selected Topic:

Slide to adjust relevance metric:(2)  
 $\lambda = 0.93$



Intertopic Distance Map (via multidimensional scaling)

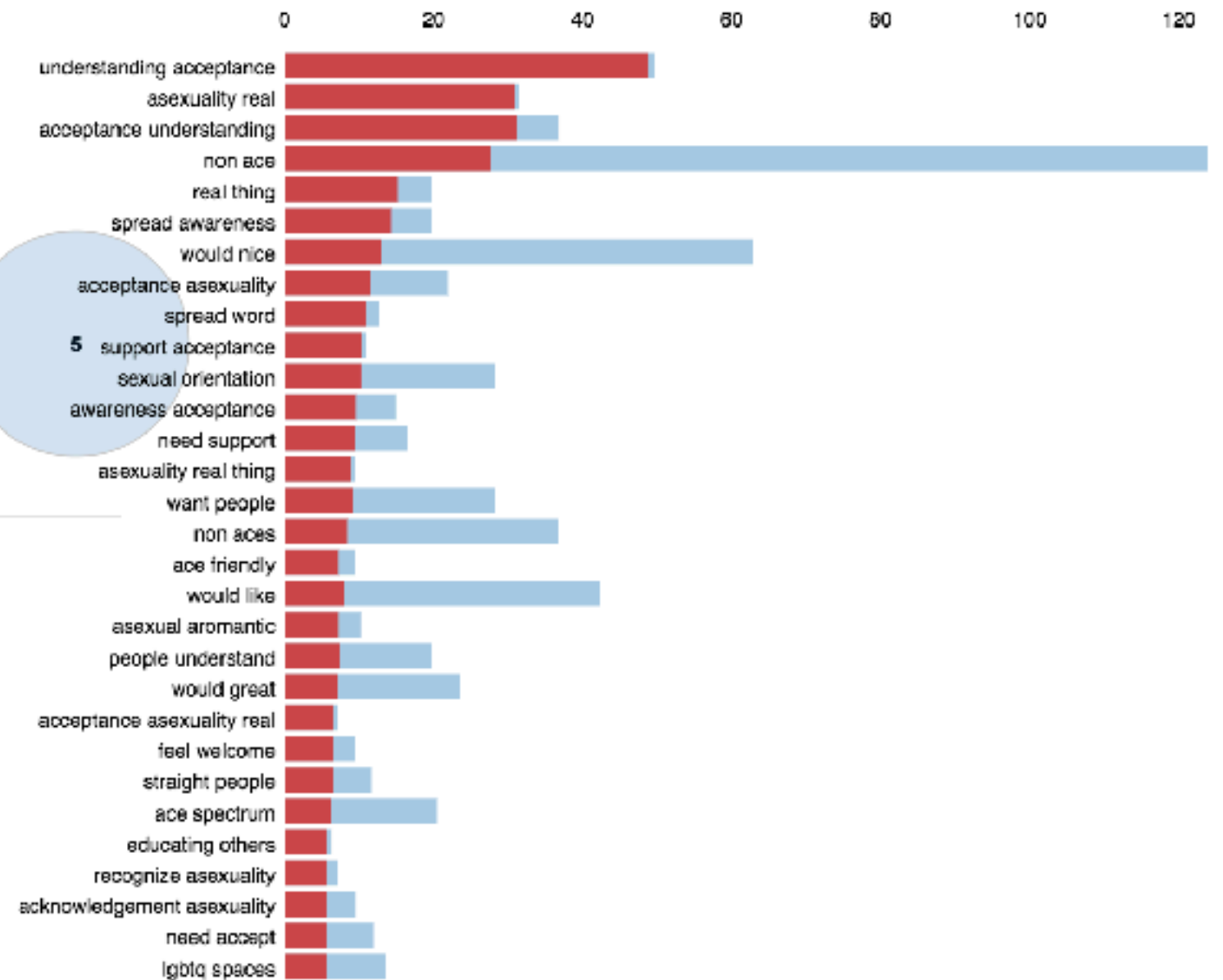
## Other Communities Cluster 2



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 2 (20.5% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

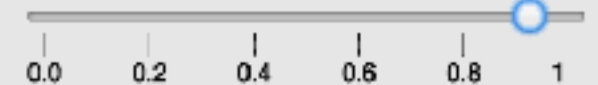
1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$  for topics  $t$ ; see Chuang et. al (2012)

2.  $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

# WHAT KIND OF SUPPORT / SERVICES DO YOU MOST NEED FROM OTHER COMMUNITIES?

Selected Topic: 2 [Previous Topic](#) [Next Topic](#) [Clear Topic](#)

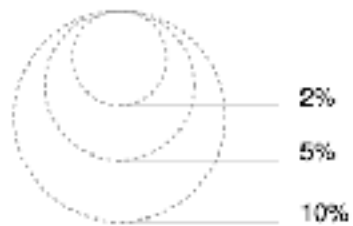
Slide to adjust relevance metric:(2)  
 $\lambda = 0.93$



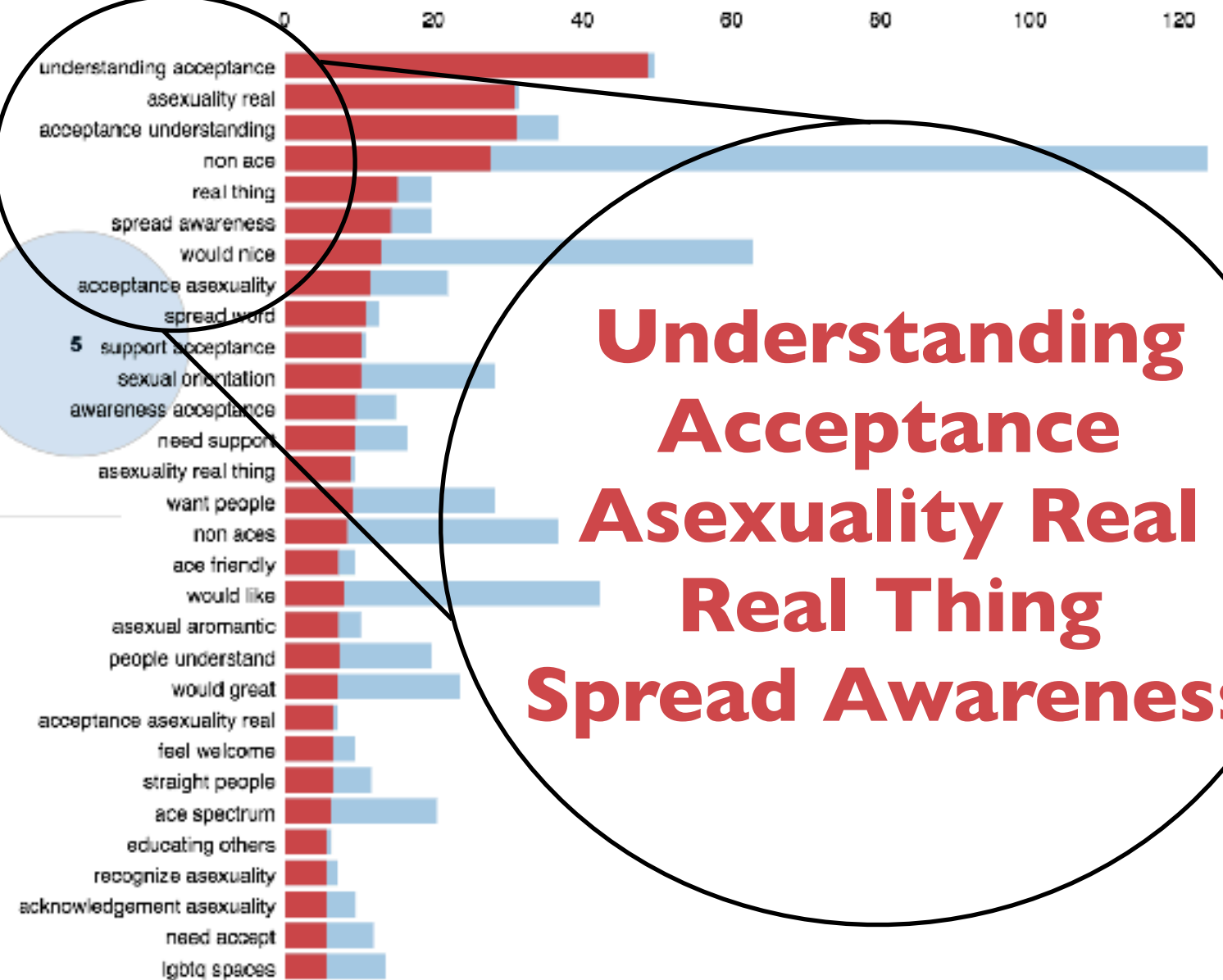
Intertopic Distance Map (via multidimensional scaling)

**Other Communities  
Cluster 2**

Marginal topic distribution



Top-30 Most Relevant Terms for Topic 2 (20.5% of tokens)



**Understanding  
Acceptance  
Asexuality Real  
Real Thing  
Spread Awareness**

Overall term frequency  
Estimated term frequency within the selected topic

1.  $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$  for topics  $t$ ; see Chuang et. al (2012)  
2.  $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Sievert & Shirley (2014)

# WHAT KIND OF SUPPORT / SERVICES DO YOU MOST NEED FROM OTHER COMMUNITIES?

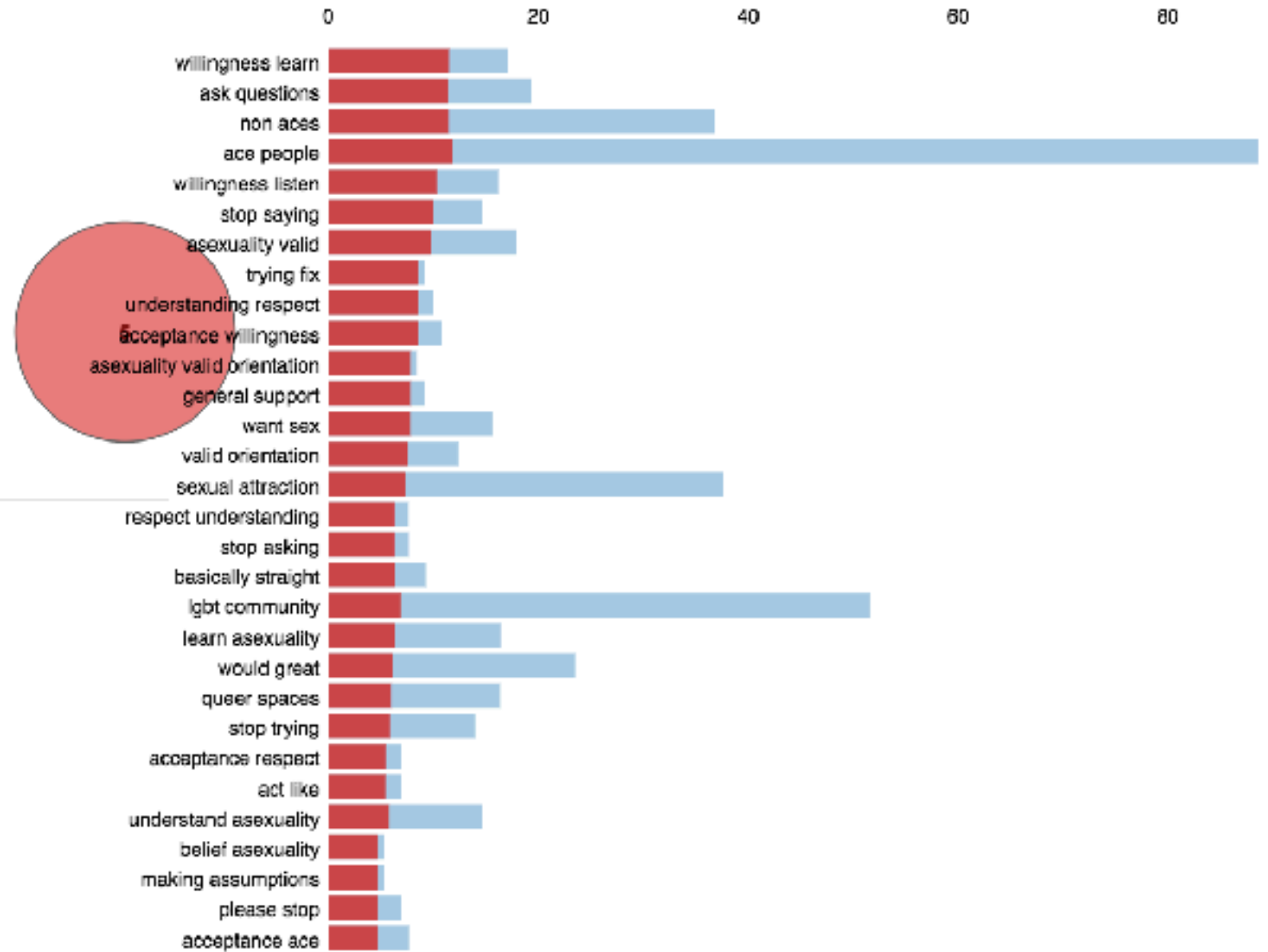
Selected Topic:

Slide to adjust relevance metric:(2)  
 $\lambda = 0.93$

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 5 (17.1% of tokens)

**Other Communities  
Cluster 5**



Marginal topic distribution



Overall term frequency  
Estimated term frequency within the selected topic

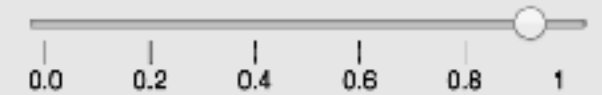
1. saliency(term  $w$ ) = frequency( $w$ ) \*  $[\sum_t p(t | w) * \log(p(t | w)/p(t))]$  for topics  $t$ ; see Chuang et. al (2012)

2. relevance(term  $w$  | topic  $t$ ) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Stevert & Shirley (2014)

# WHAT KIND OF SUPPORT / SERVICES DO YOU MOST NEED FROM OTHER COMMUNITIES?

Selected Topic: 5   Previous Topic   **Next Topic**   Clear Topic

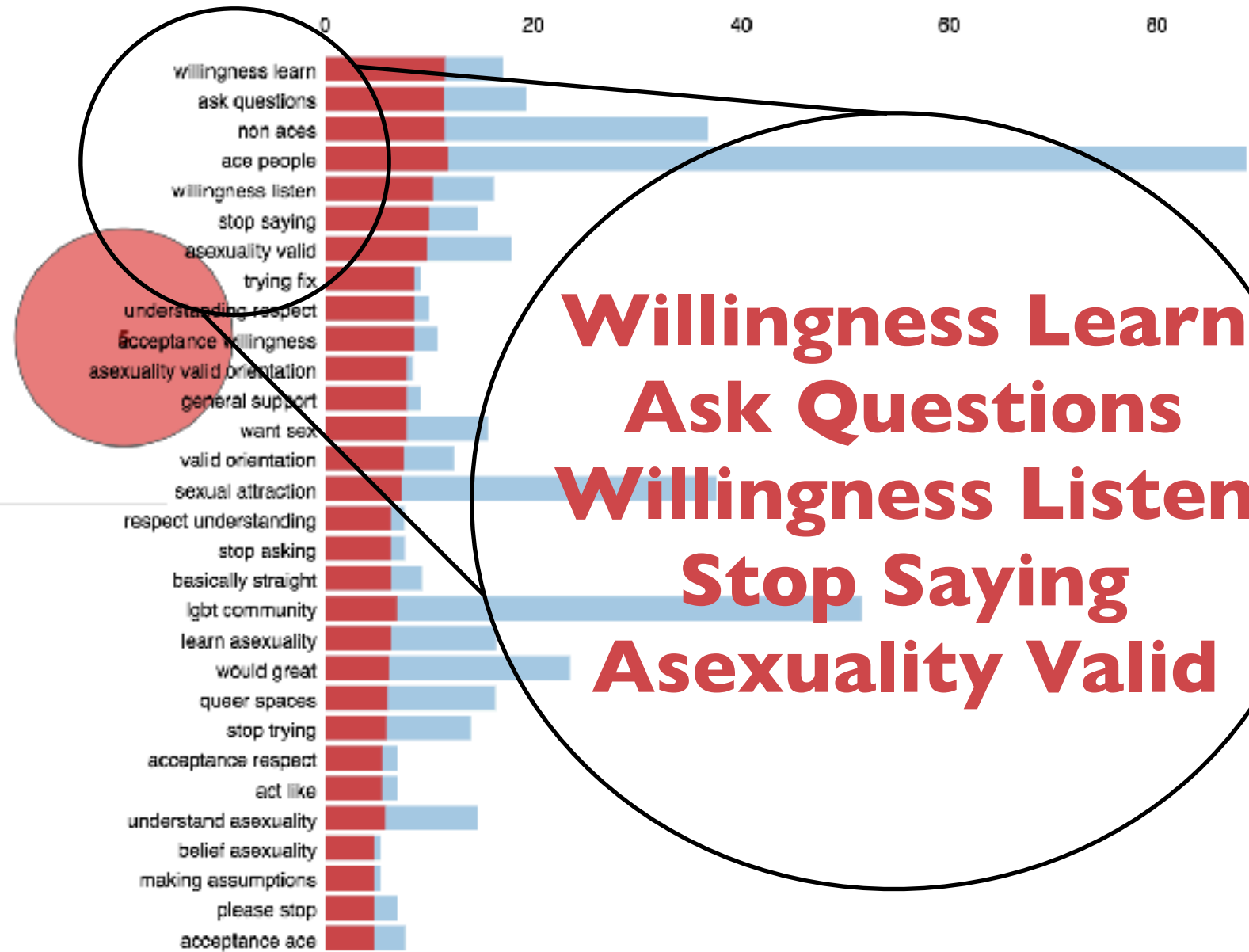
Slide to adjust relevance metric:(2)  
 $\lambda = 0.93$



Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 5 (17.1% of tokens)

**Other Communities  
Cluster 5**



**Willingness Learn  
Ask Questions  
Willingness Listen  
Stop Saying  
Asexuality Valid**

Marginal topic distribution



Overall term frequency  
Estimated term frequency within the selected topic

1. saliency(term  $w$ ) = frequency( $w$ ) \* [sum  $t$   $p(t | w) * \log(p(t | w)/p(t))$ ] for topics  $t$ ; see Chuang et. al (2012)  
2. relevance(term  $w$  | topic  $t$ ) =  $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$ ; see Stevert & Shirley (2014)

# TOPICS



# DISCRIMINATION









# DISCRIMINATION

# TOPICS

# Feeling Broken

## Sexual Jokes



*Girlfriend Insisted*

*Suggestions  
Fix Asexuality*

*Feeling Broken*

*Sexual Jokes*



# TOPICS



# Ace Lookup

Lookup Word:

A large, empty gray rectangular box with a thin black border, intended for displaying the lookup results. In the bottom right corner, there is a small icon consisting of two parallel diagonal lines.



# SUMMARY

- Can use **topic modeling to analyze** long form text answers
- **Reduce manual work** by volunteers analyzing the survey
- Allows **community members** to **investigate** and interact with the dictionary for themselves

# FUTURE

- Create interactive graphs for descriptive statistics
- Filter for certain populations



Questions?

# LONG FORM QUESTIONS

- Do you tell other people about your asexuality? How have they responded, and how did you feel about those responses?
- Have you ever experienced prejudice or discrimination as a result of your a/sexuality? What happened?
- What kind of support or services do you most need from ace communities?'
- What kind of support or services do you most need from non-ace allies?
- What unanswered questions do you have about the demographics, behaviors, or attitudes of aces and ace communities?
- Are there any other stories or statements you would like to share?