# Some stuff on identities and networks and stereotypes and text

**Kenny Joseph**
kjoseph@cs.cmu.edu

Carnegie Mellon

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

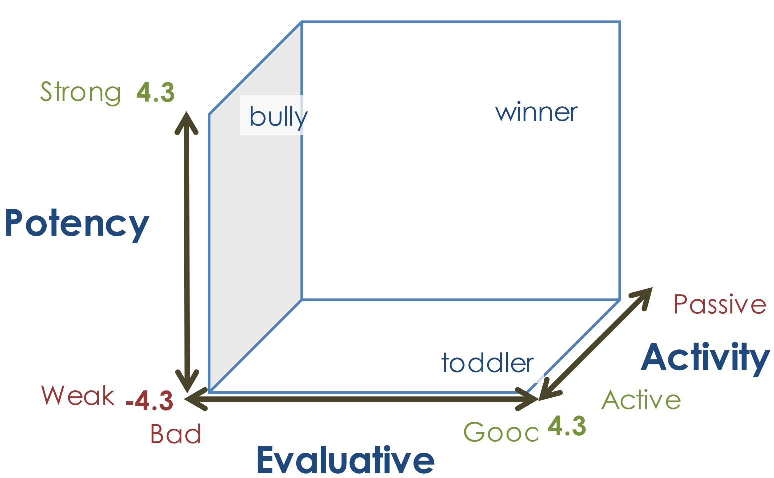
**Identities** are the **words and phrases** we use to **label other people**

2

**Stereotypes** are the **meanings** conveyed by an identity



An EPA profile is a position in this 3D space

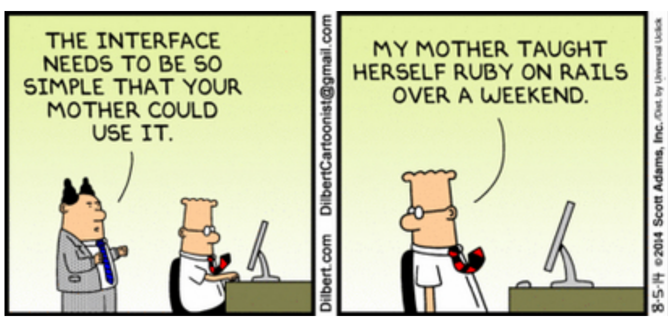



**Affective stereotypes** are defined by how we **feel** about identities

**(Osgood, 1969; Heise 1987 )**


**Semantic stereotypes** refer to **relationships we presume between identities**

5




Our identities and the stereotypes they carry have important effects on our lives

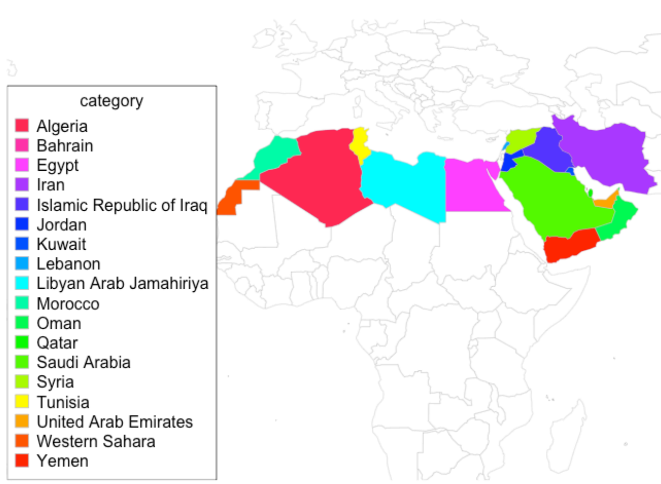
6

# Overview

- Extracting affective stereotypes using "social event networks"
- Extracting a network model of stereotypes
- Networks of identities



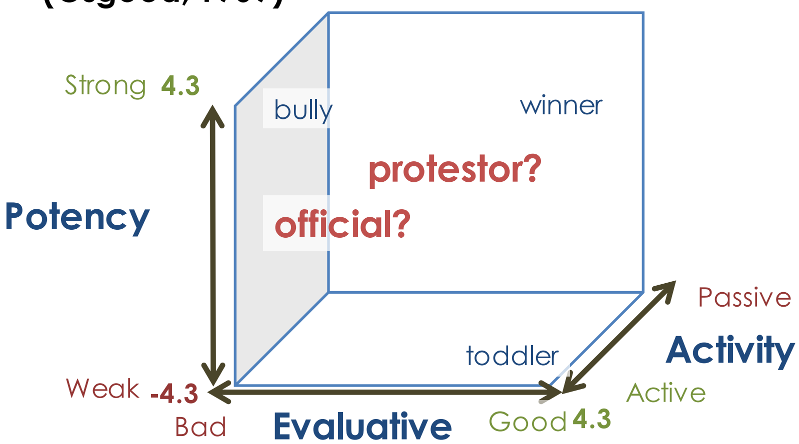
# The Data




- Newspaper data
  - 600K articles
  - LexisNexis, centered on 16 MENA countries
  - Major news outlets
  - 7/10 – 12/12

# Measuring Stereotypes with ACT

(Osgood, 1969)




- An **affective, attributional** measurement model



# Inferring Stereotypes using ACT

? officials

− criticize

+ women

ACT gives a mathematical model for how **social events imply stereotypes**

## Caveat to applying event model

? officials

− accused ?

? protestors



## Soln. – allow stereotypes to "diffuse"

| | | | |
|---|---|---|---|
| ? | officials | ? | officials |
| − | criticize | − | accused |
| + | women | ? | protestors |

# A caveat to applying event model

Sometimes,

officials ≠ officials

Solution:

Assume multiple latent stereotypes of each identity/behavior exist



# An overview of the approach

# More on extracting events, identities

1. Ran dependency parser, extracted all N -> V -> N
2. Cleaned text using, e.g., stemming (accused -> accuse)
3. Hand-curated list of identities and behaviors

- 102 identities, 87 behaviors, 10K events
- Only 44% of identities in ACT dicts



# The Statistical Model

$$\pi \sim \mathrm{Dirichlet}(p)$$
$$\mu_0 \sim \mathrm{N}(m_0, \frac{\sigma_0^2}{k_0})$$
$$\sigma_0^2 \sim \mathrm{Inv}{-}\chi^2(v_0, s_0)$$

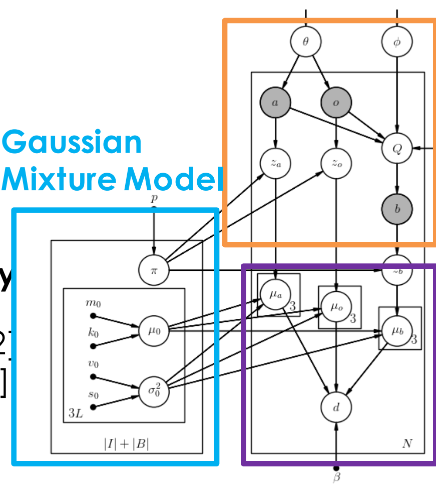
$$\phi \sim \mathrm{Dirichlet}(\psi)$$
$$\theta \sim \mathrm{Dirichlet}(\alpha)$$
$$a \sim \mathrm{Categorical}(\theta)$$
$$o \sim \mathrm{Categorical}(\theta)$$
$$b \sim \mathrm{Categorical}(Q)$$

$$z_a \sim \mathrm{Categorical}(\pi) \quad z_b \sim \mathrm{Categorical}(\pi) \quad z_o \sim \mathrm{Categorical}(\pi)$$
$$\mu_a \sim \mathrm{N}(\mu_{0,z_a}, \sigma^2_{0,z_a}) \quad \mu_b \sim \mathrm{N}(\mu_{0,z_b}, \sigma^2_{0,z_b}) \quad \mu_o \sim \mathrm{N}(\mu_{0,z_o}, \sigma^2_{0,z_o})$$
$$d \sim \mathrm{Laplace}(\sum_i^9 (f_i - M_{i*}^T G(f))^2, \beta) \quad \text{where} \quad f = [\mu_{a_e}, \mu_{a_p}, \mu_{a_a}, \mu_{b_e}, \mu_{b_p}, \mu_{b_a}, \mu_{o_e}, \mu_{o_p}, \mu_{o_a}]$$

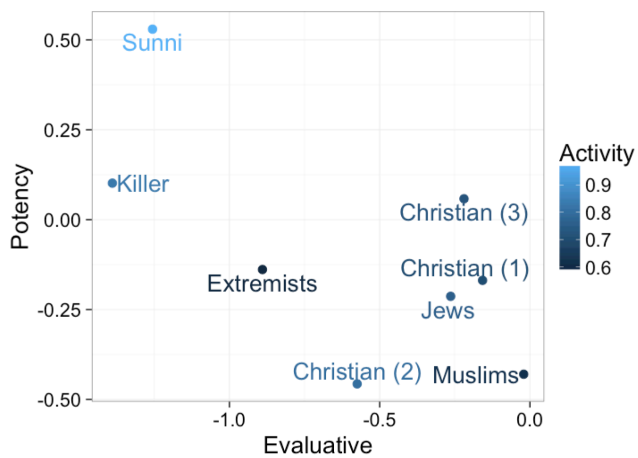
**Officials criticize demonstrators**

**ACT Dictionary**

criticize [-4 3 2]
women [4 1 4]
....



# One Result w.r.t. religious identities

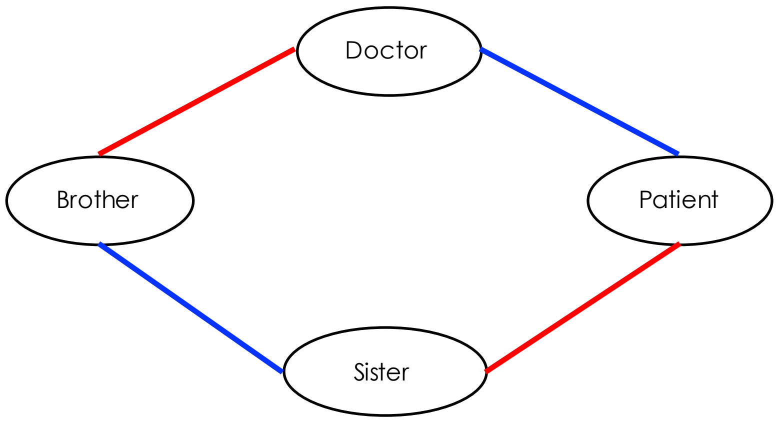



- Sunnis universally bad, powerful
- Explanation:
  - Events on the ground
  - Western media bias?



## NETWORK MODELS OF STEREOTYPE

# Parallel Constraint Satisfaction Models

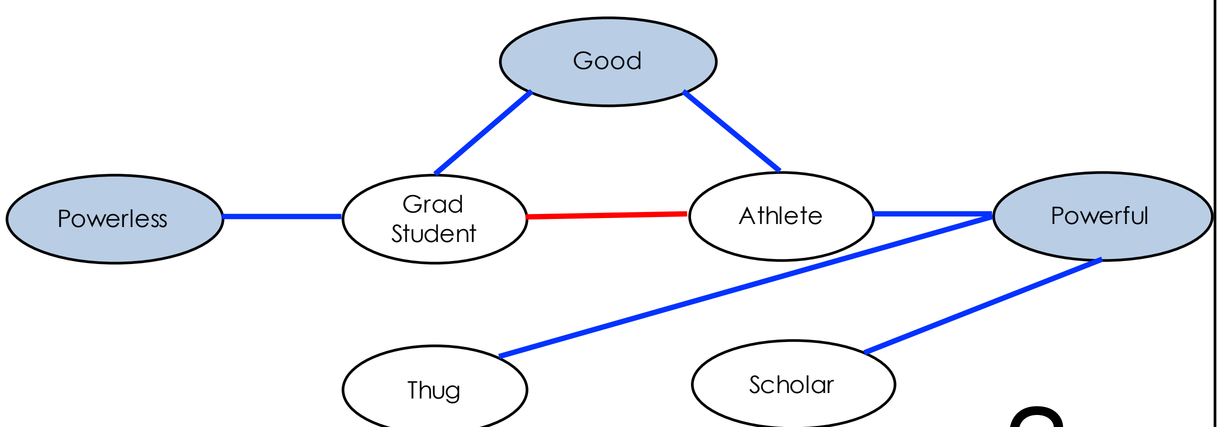



Links in PCSMs define semantic stereotypes

PCSMs are essentially Markov Random Fields through which *cognitive activation* flows

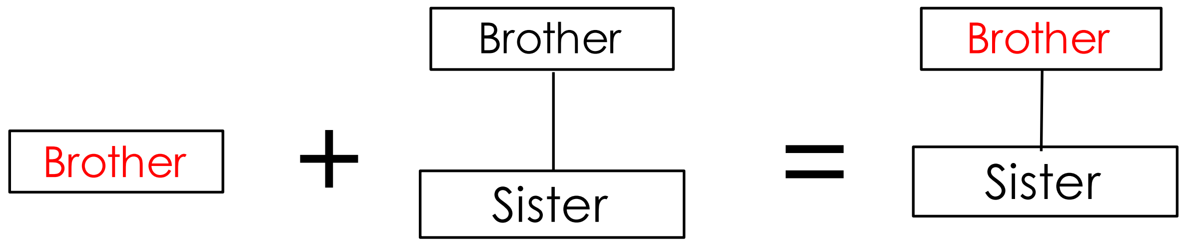


# Hard to model Affect in PCSMs




?

## Combining existing models

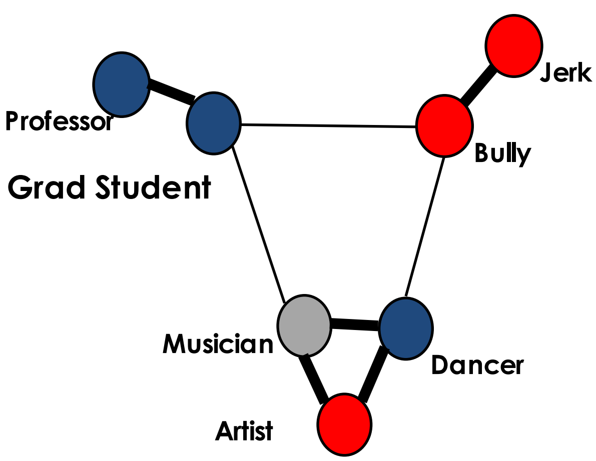



**Attributional**
Parsimonious, Affective
No semantic relationships

**Relational**
Cognitively More Plausible, Semantics
No affective meaning



## Affective + Semantic Network of Stereotypes




Stereotypes as an **attributed network**

Now, how do we "learn" from Twitter data?

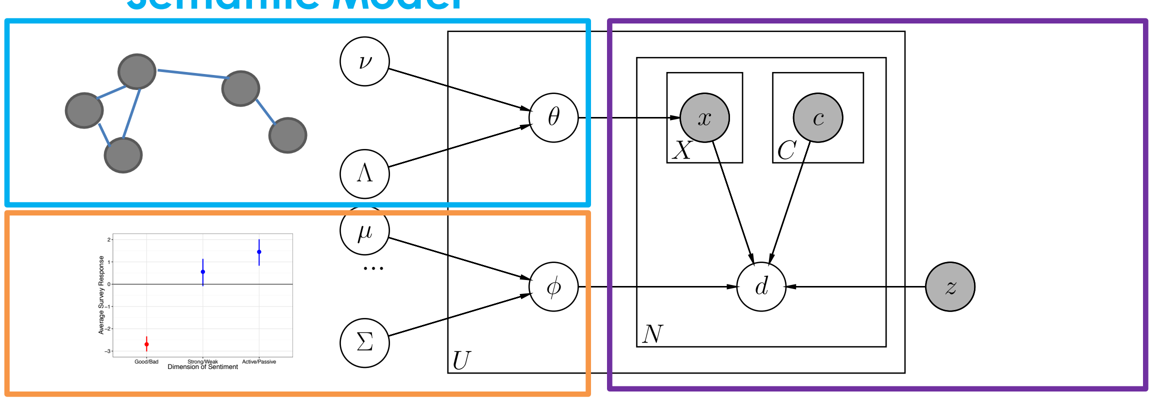
# Data Used (Population considered)

- Twitter data
  - Subset of 50K users from Study 2
  - Subsetting based on more restricted bot/celeb removal, gender tagging (gender not used)
- 310 identities of interest
  - From popular identities in Study 2 results; some domain relevant
- Sentiment data (EPA profiles)
  - Smith-Lovin et al. (2015)
  - Warriner et al. (2014)



# A Statistical Model for Twitter Data

Semantic Model




Affective Model (Survey data priors)

Tweet data (per user)

## Generative model – affective stereotypes

$$p(\mu, \Sigma) \sim \mathcal{NIW}(\mu_0, \Sigma_0, \kappa_{0,S}, \gamma_{0,S})$$
$$p(\phi) \sim \mathcal{N}(\mu, \Sigma)$$
$$p(d) \sim \text{Laplace}(q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z), \beta)$$

- Draw per-identity distrib. in EPA space from survey priors
- Draw per-user EPA profiles from this distribution
- Draw per-tweet "deflection" balancing by user's current views, constraints in tweet



## Details on deflection

$$p(d) \sim \text{Laplace}(q_{u,n}(\phi_u, X_{u,n}, C_{u,n}, z), \beta)$$

- In ACT, deflection defines likelihood of social event
  - "Teacher instructs student" has low deflection
  - "Teacher hits student" has high deflection
- I use the same concept for likelihood of a **tweet**
- Like social event "suggests", or **constrains**, EPA profiles for identities, so too does **text in a tweet**
- Formalize using **quadratic constraints**, like ACT does for event model

## Strategy for mining affect

So terrible that a young man was killed by a police officer.

For each identity of interest:

- Identify any **social events** it is involved in
  - man (young) -> killed_by -> police_officer
- Find any "**sentiment words**" (in our sentiment dictionary) in the tweet
- Construct **q** by summing constraints –
- "Terrible" constraint on police officer ($\phi_{po}$):

$$(\phi_{po,e} - ter_e)^2 + (\phi_{po,p} - ter_p)^2 + (\phi_{po,a} - ter_a)^2$$

Identity of interest

Other identity

Verb we have sentiment info for

Modifier we have sentiment info for



## Validation – Semantic model

| Associative Model | Ppl. |
|---|---|
| Simple | 4.864 |
| User Baseline | 4.474 |
| **Our Model** | **4.363** |

**Metric: Perplexity** of identities in left-out data (**lower is better)**

**Fill in the blank (on left out data):**
___ rule, boys drool

Baselines:
- **Simple:** Just based on frequency of each identity
- **User:** Laplace-smoothed language model

# Validation – Affective model

| Affective Model | Avg. Rank |
|---|---|
| Simple | 134.744 |
| User Baseline | 127.272 |
| **Our Model** | **126.042** |

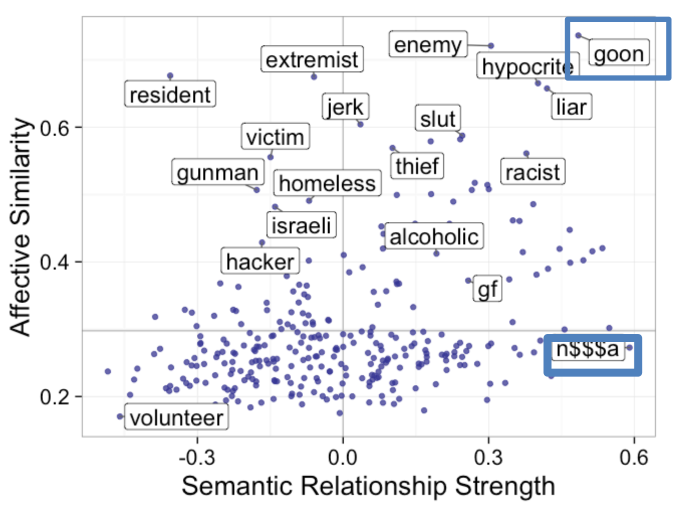
**Metric: average rankings** of identities in left-out data (**lower is better)**

**Fill in the blank (on left out data):**
___ rule, boys drool

- Baselines:
  - **Simple:** Tweet-based average using VADER
  - **User:** Simple back-off tweet-based model using VADER



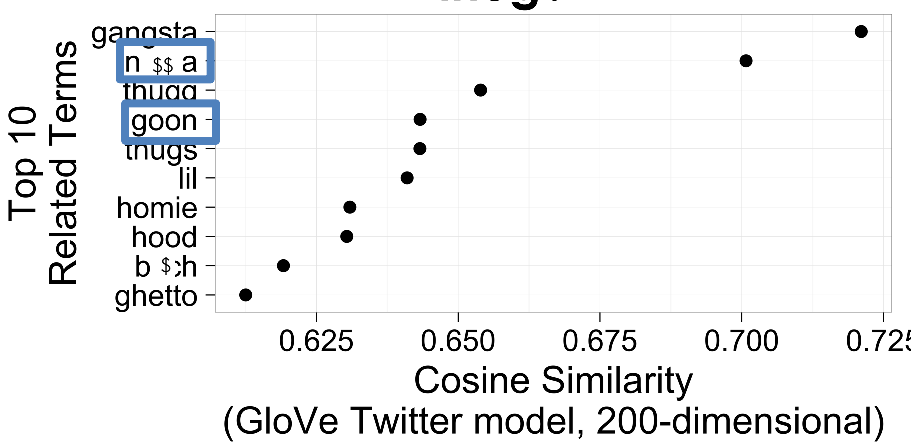
# Results for Thug




- Top right – affectively similar & semantically related
- N(-a) word semantically, not affectively related

# Existing NLP methods - Thug

**E.g. deep learning… what words are related to thug?**





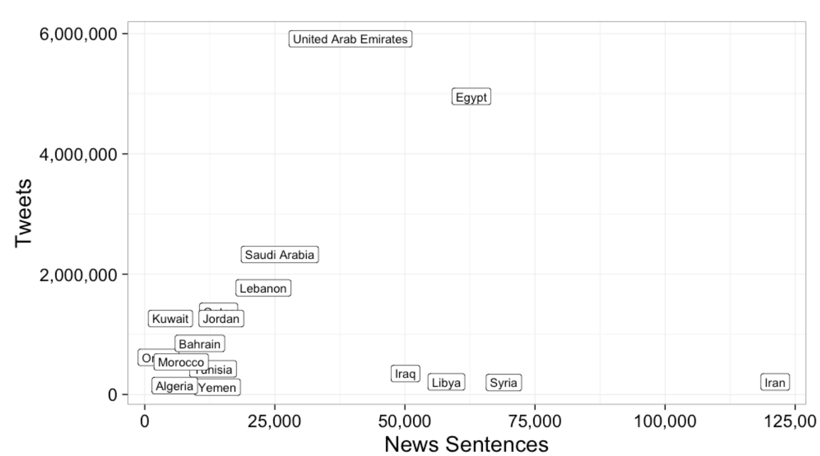

## NETWORKS OF IDENTITIES

# Approach

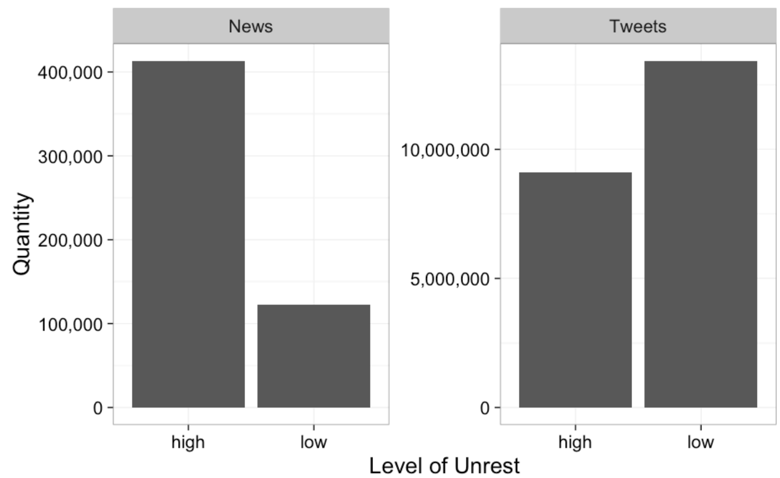
- Twitter data
  - 150K Twitter users who sent >5 tweets from within the original Arab Spring dataset
- News data
  - Original news data
- Construct common vocabulary; common data format
- Run through Bamman et al. Word2Vec embedding model
- Determine list of interesting identities
  - 280 identities prevalent in both datasets
- Construct network of similarity between these identities for High/Low stability, News/Twitter (4 networks total)



# Tweets/News Sentences count by country



# Tweets/News articles by unrest level of country

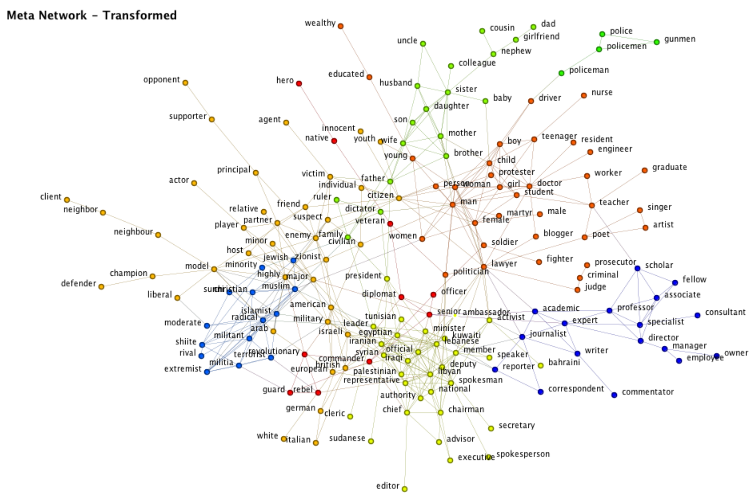





# High/Low Civil Unrest Categorization

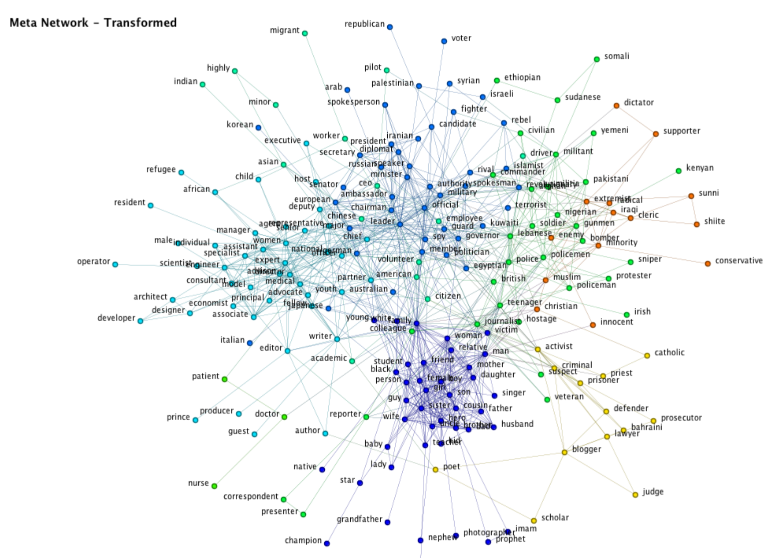
| High Unrest Countries | Low Unrest Countries |
|---|---|
| Bahrain | Qatar |
| Iraq | Kuwait |
| Iran | Morocco |
| Libya | Jordan |
| Algeria | Saudi Arabia |
| Egypt | Oman |
| Syria | United Arab Emirates |
| Tunisia | Yemen |
| Lebanon | |

# High Unrest, Newspaper data (.7 cutoff, LCC)

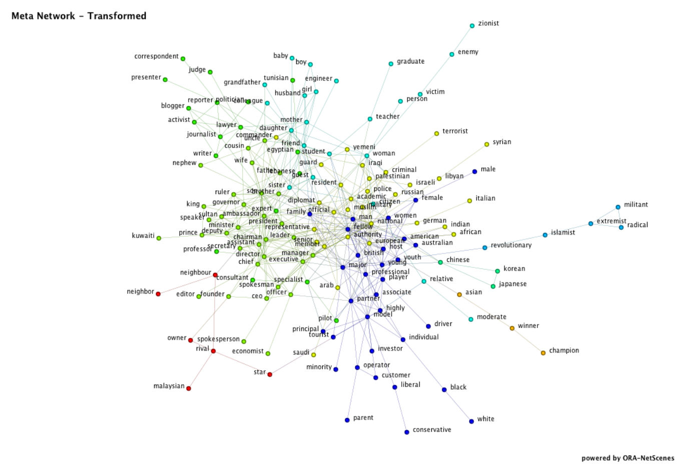





# High Unrest, Twitter (.65 cutoff, LCC)



# Low Unrest, News (.73, LCC)






# Low Unrest, Twitter (.65 cutoff)

# Conclusion

- Extracting affective stereotypes using "social event networks"
- Extracting a network model of stereotypes
- Networks of identities

- Many different ways to think about identities, text and networks!