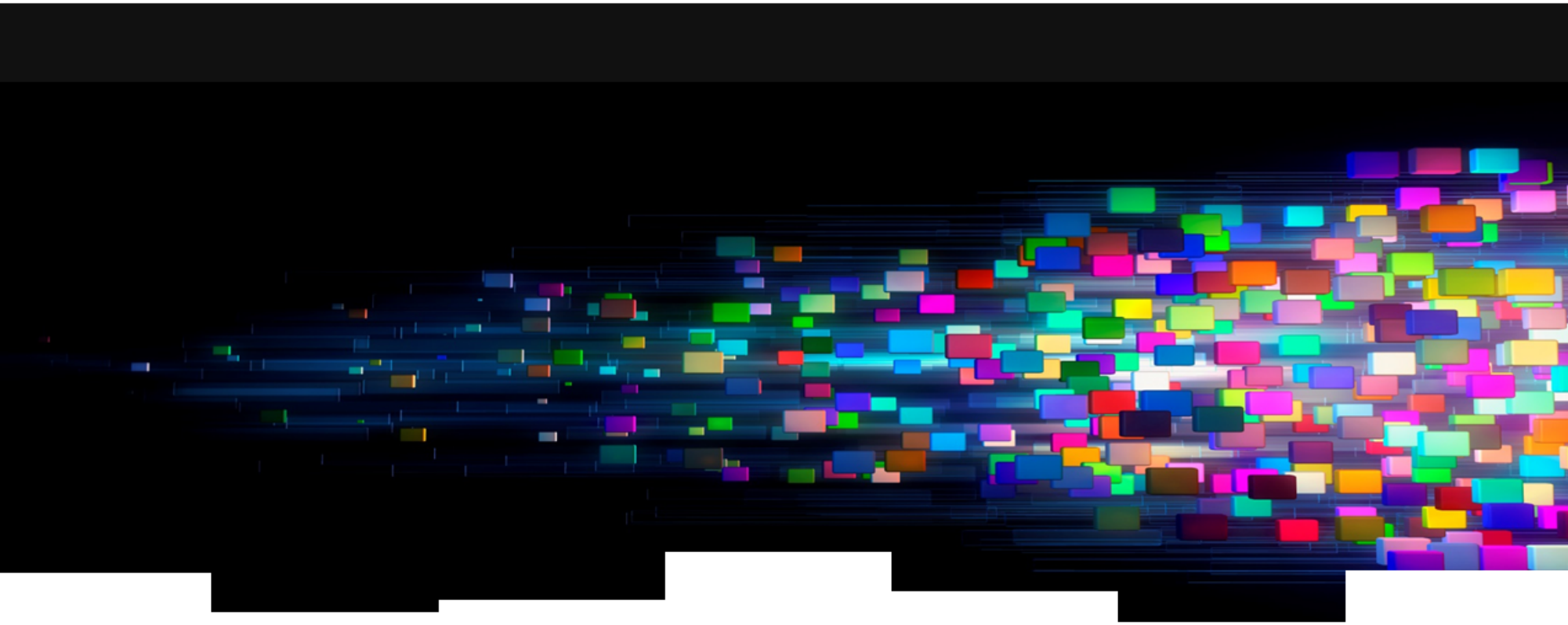
Detection of latent roles in online forums

Luchon- July 1, 2014

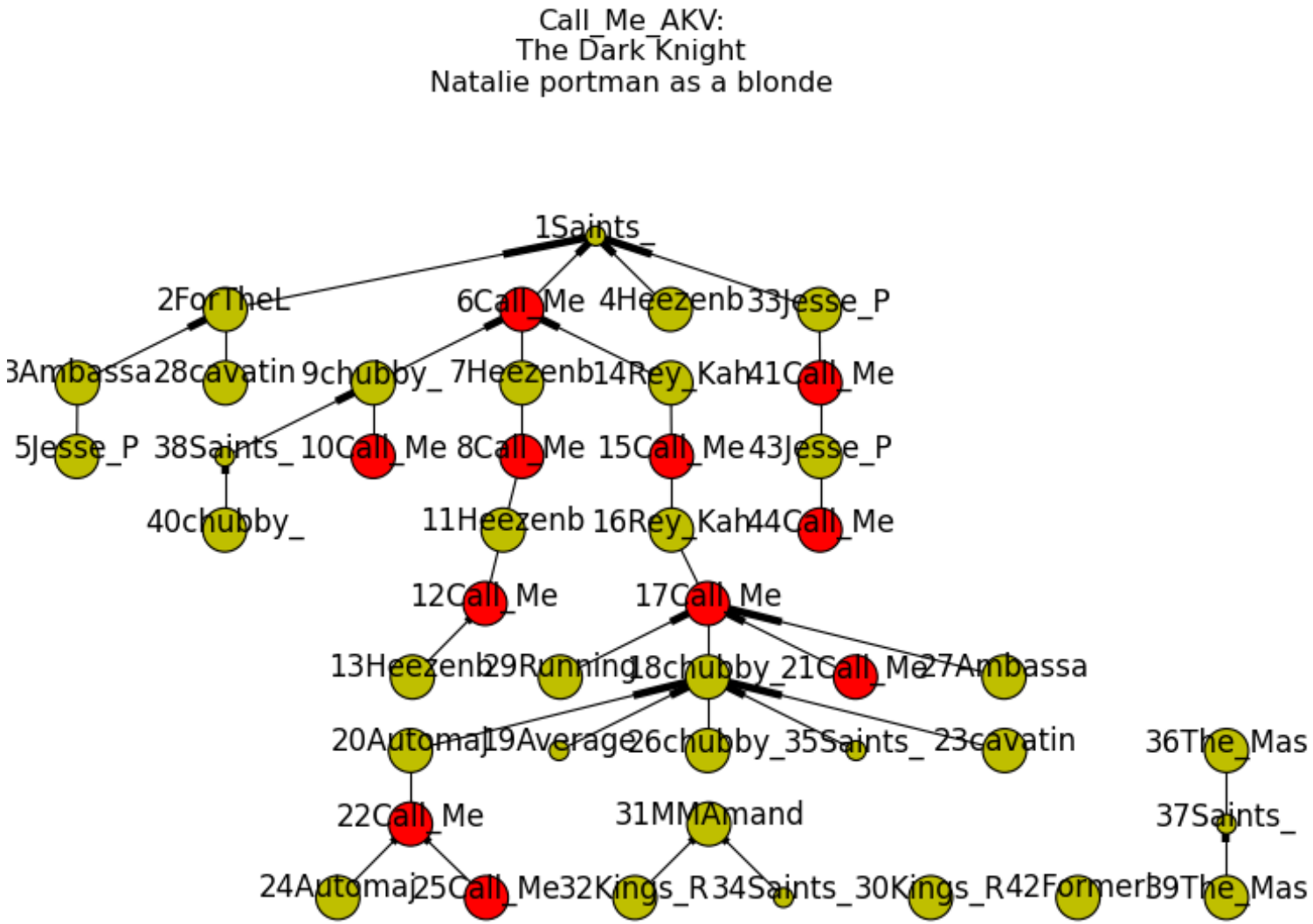


Alberto Lumbreras,
Sup: Jouve B., Velcin J.

eric
Entrepôts, Représentation & Ingénierie des Connaissances

technicolor

Roles in discusion threads



- Task: detect roles
- Definition: role as archetypical behavior or social function.

technicolor

Different roles, different definitions.

Sociology/Antropology

- **attributes**: strategies of speech.
- **technique**: ethnology, observational study.
- **Identified roles**: Celebrity, Newbie, Lurker, Flamer, Troll, Ranter.

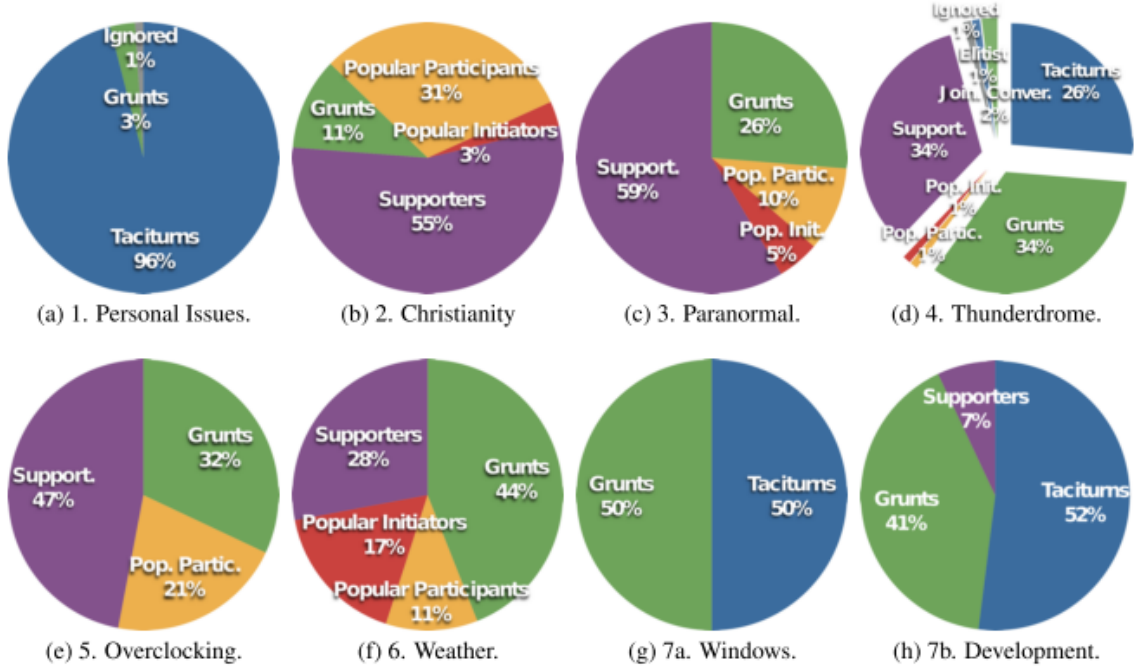
Cartoon by Maurice Henry, published in La Quinzaine Litteraire, 1 July 1967.

[1] S. Golder and J. Donath, "Social roles in electronic communities," Internet Res., vol. 5, 2004.

technicolor

Similar attributes

- **attributes**: in-deg, out-deg, %init, %posts replied, % bi-dir neighs,...
- **technique**: clustering.
- **Identified roles**: Joining conversationalists, Popular initiators, Taciturns, Supporters, Elitists, Popular participants, Grunts, Ignored.



[2] J. Chan, C. Hayes, and E. Daly, "Decomposing discussion forums using common user roles," in Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, 2010.

technicolor

Similar relationships

- **attributes**: sociomatrix (matrix of relations)
- **technique**: blockmodeling.
- **Identified roles**: Centre-periphery, hierarchies, horizontal structures, ghettos...

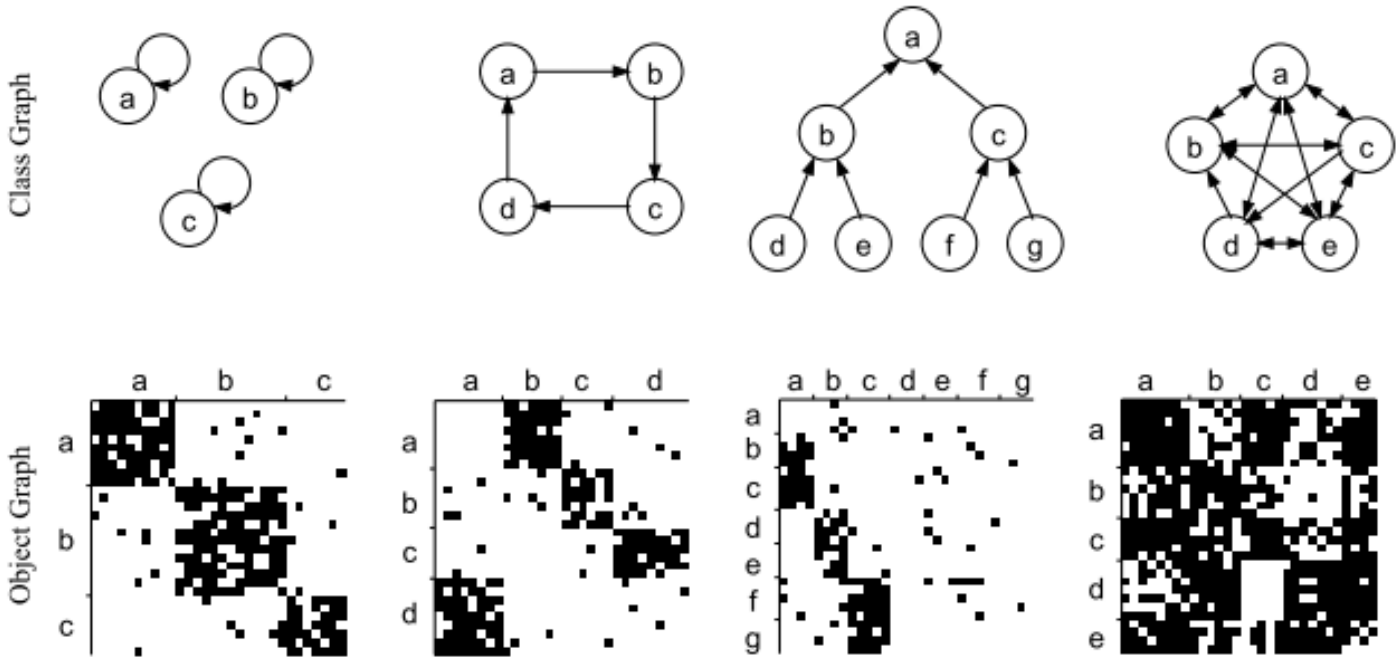


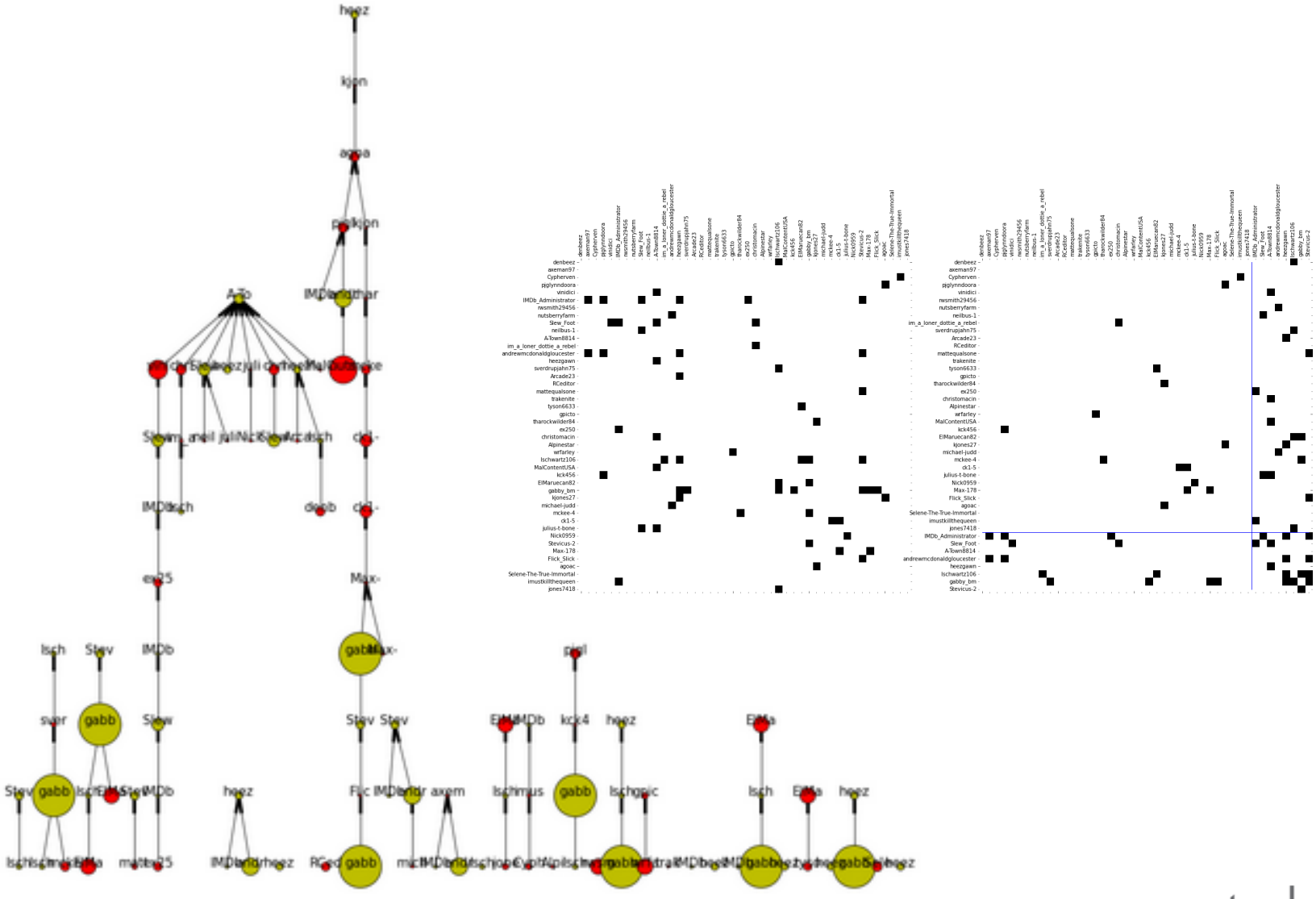
Figure: Kemp, C., Griffiths, T. & Tenenbaum, J., 2004. Discovering latent classes in relational data.

[1] H. White, S. Boorman, and R. Breiger, "Social structure from multiple networks. I. Blockmodels of roles and positions," Am. J. Sociol., 1976.
[2] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," J. Am. Stat. Assoc, 2001.

technicolor

Similar relationships

Example



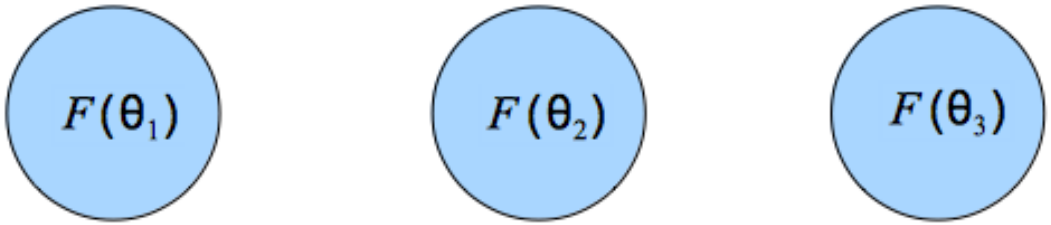
Role as similar behavior

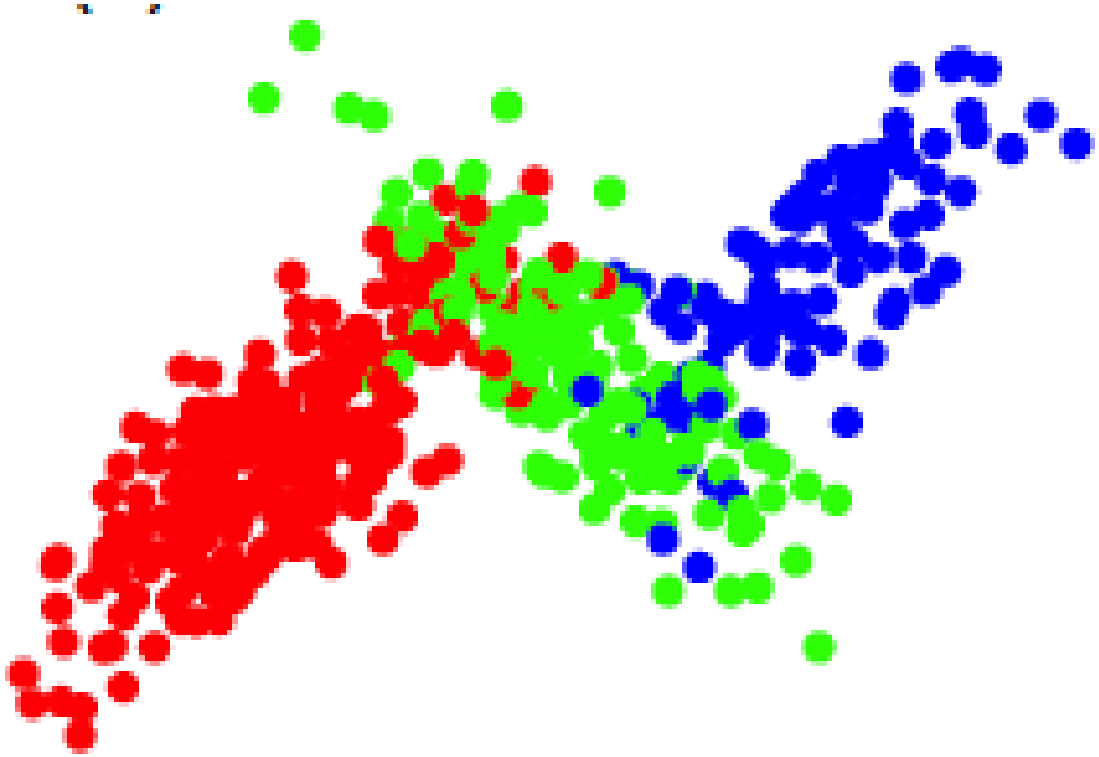
Idea: if you hold role r, you behave like the archetype r plus some noise.

$$b_u = r_u + \epsilon_u \tag{1}$$

$$\text{(toy example) } b_u \sim \mathcal{N}(r_u, \epsilon_u) \tag{2}$$

Intuition





technicolor

Bayesian framework

Bayesian probability:

$$\underbrace{P(\theta|Y)}_{posterior} = \frac{\overbrace{P(Y,\theta)}^{\text{joint probability}}}{\int_\theta P(Y,\theta)} = \frac{\overbrace{P(Y|\theta)}^{likelihood}\overbrace{P(\theta)}^{prior}}{\int_\theta P(Y|\theta)P(\theta)} \propto \overbrace{P(Y|\theta)}^{likelihood}\overbrace{P(\theta)}^{prior} \tag{3}$$

BAYESIAN BONUS: we can make predictions (and therefore validate our model).

$$P(y|y_{t-1},\theta) \tag{4}$$

Mixture models

A generative story:

$$behavior_u | role_u, \theta_{role} \sim F(behavior | role_u, \theta_{role}) \tag{5}$$

$$\theta_{role} | \beta \sim G(\beta) \tag{6}$$

$$role_u \sim Discrete(P(role_1), ..., P(role_K)) \tag{7}$$

$$P(role_1), ..., P(role_K) | \alpha \sim Dirichlet(\alpha) \tag{8}$$

(intuition: imagine F is a Normal distribution, role is the mean μ, and behavior is the observation y)
Probability of everything:

$$P(\mathbf{b}, \mathbf{r}, \pi, \theta) = P(\pi|\alpha) \prod_U P(r_u|\pi) \prod_K P(\theta_r|\beta) \prod_U P(b_u|r_u, \theta_{r_u}) \tag{9}$$

Intractable: Marginal probability of r:

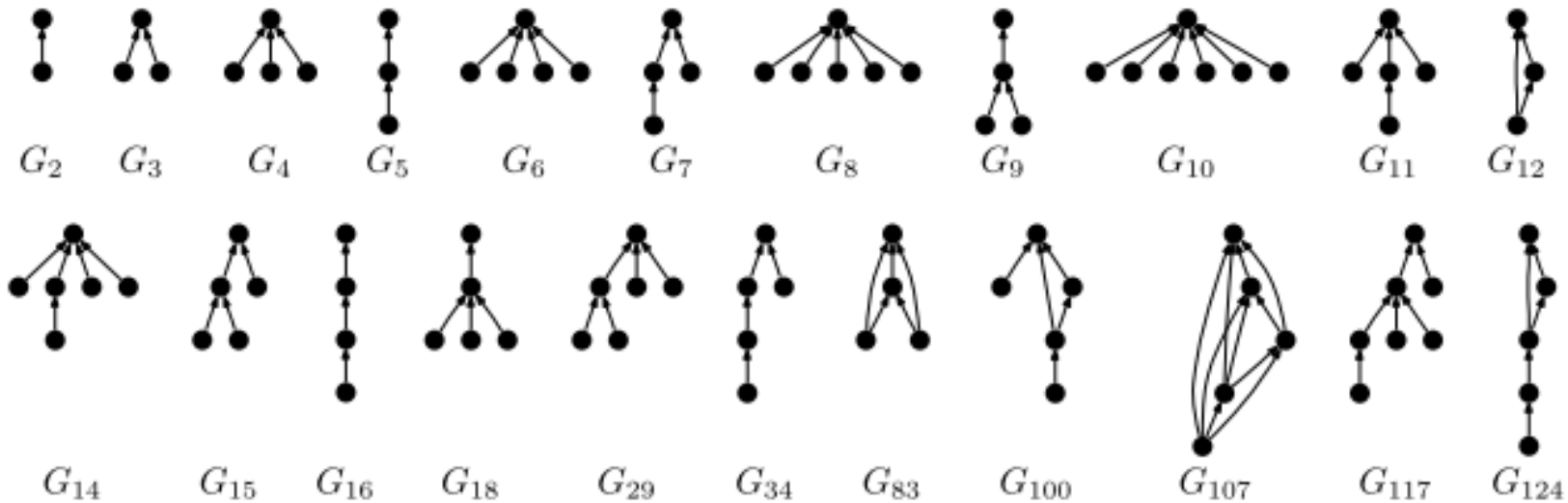
$$P(\mathbf{r}) = \sum_b \sum_\pi \sum_\theta P(\mathbf{b}, \mathbf{r}, \pi, \theta) \tag{10}$$

Solution: Gibbs sampling:

1. Loop: (11)
   $$r_u \sim P(r_u | r_{-u}, \theta, \mathbf{b}) \tag{12}$$
   $$\theta_k \sim P(\theta_k | \theta_{-k}, r, \mathbf{b}) \tag{13}$$
   $$\pi \sim P(\pi | \theta, r, \mathbf{b}) \tag{14}$$
2. Histogram r_u (15)

Behaviors

- Triads in which user is seen.
- Cascades after user participation.



Leskovec et al, "Cascading Behavior in Large Blog Graphs Patterns and a model."

- Preference function (patterns of choices).
- etc.

Remarks

- Mixture models as natural framework to group fuzzy behaviors.
- Flexibility in what behaviors to study. (structural, text, dynamics...)
- The main issue: inference (sampling)

Machine Learning:

- Non-parametric model (let the data speak)
- Efficient sampling methods (parallel, hamiltonian monte carlo...)

Thanks!