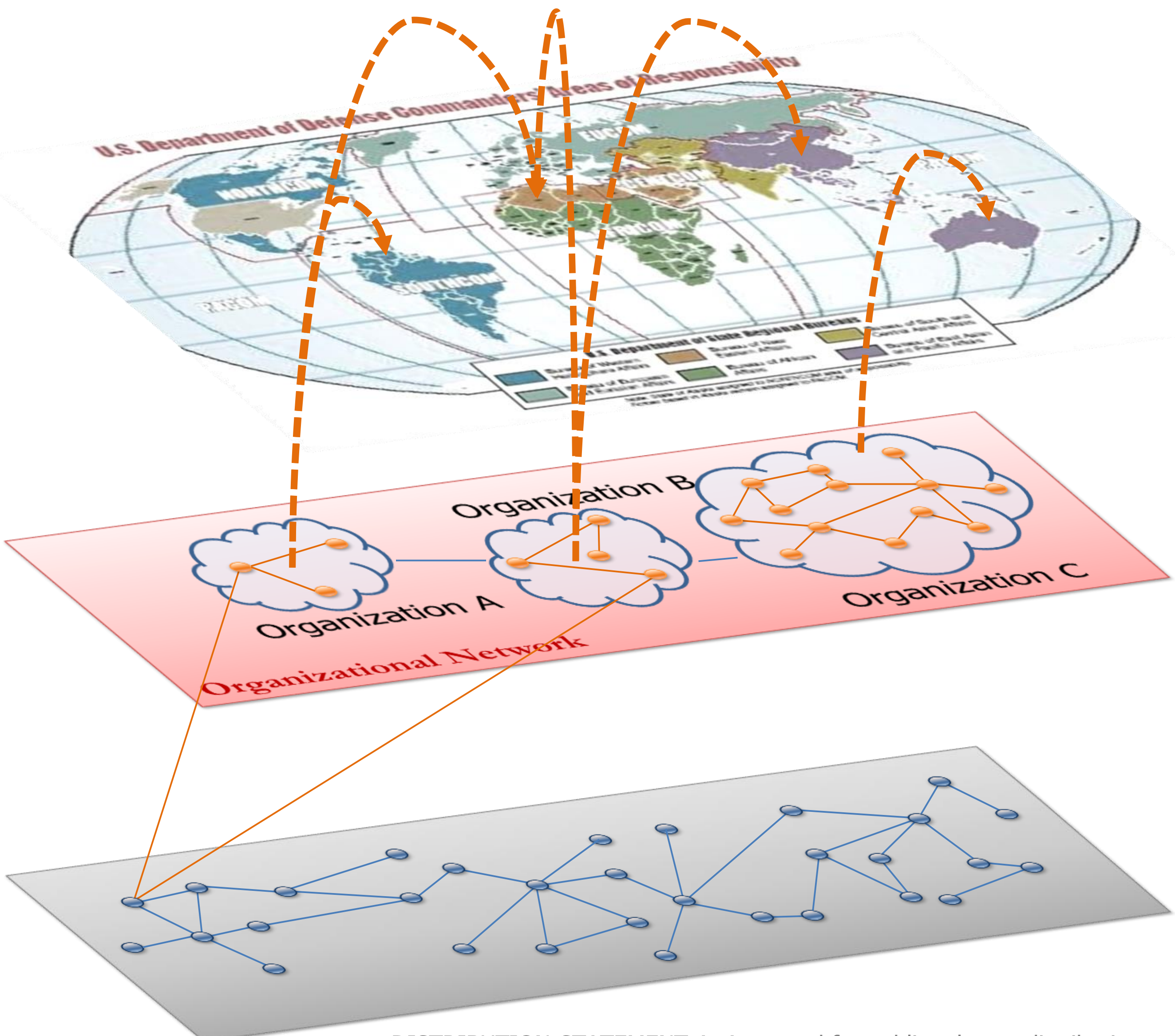
Data Science for Organizational Modeling

Understanding Organizations and Their Relationships



Traditional DOD (Nation State)

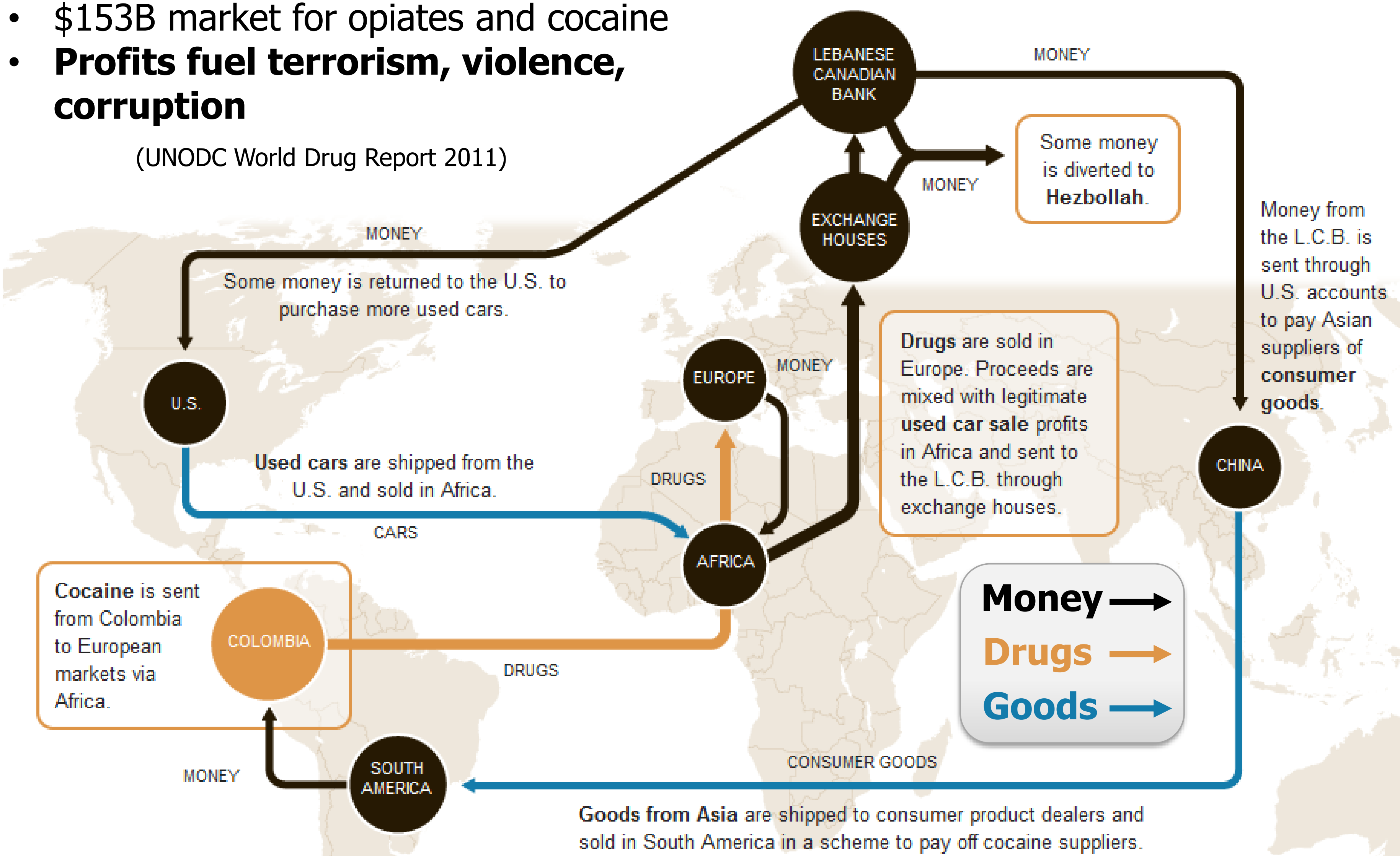
Unexplored Territory

Historic Military Counter-Insurgency

Transnational Networks with Camouflaged Participants

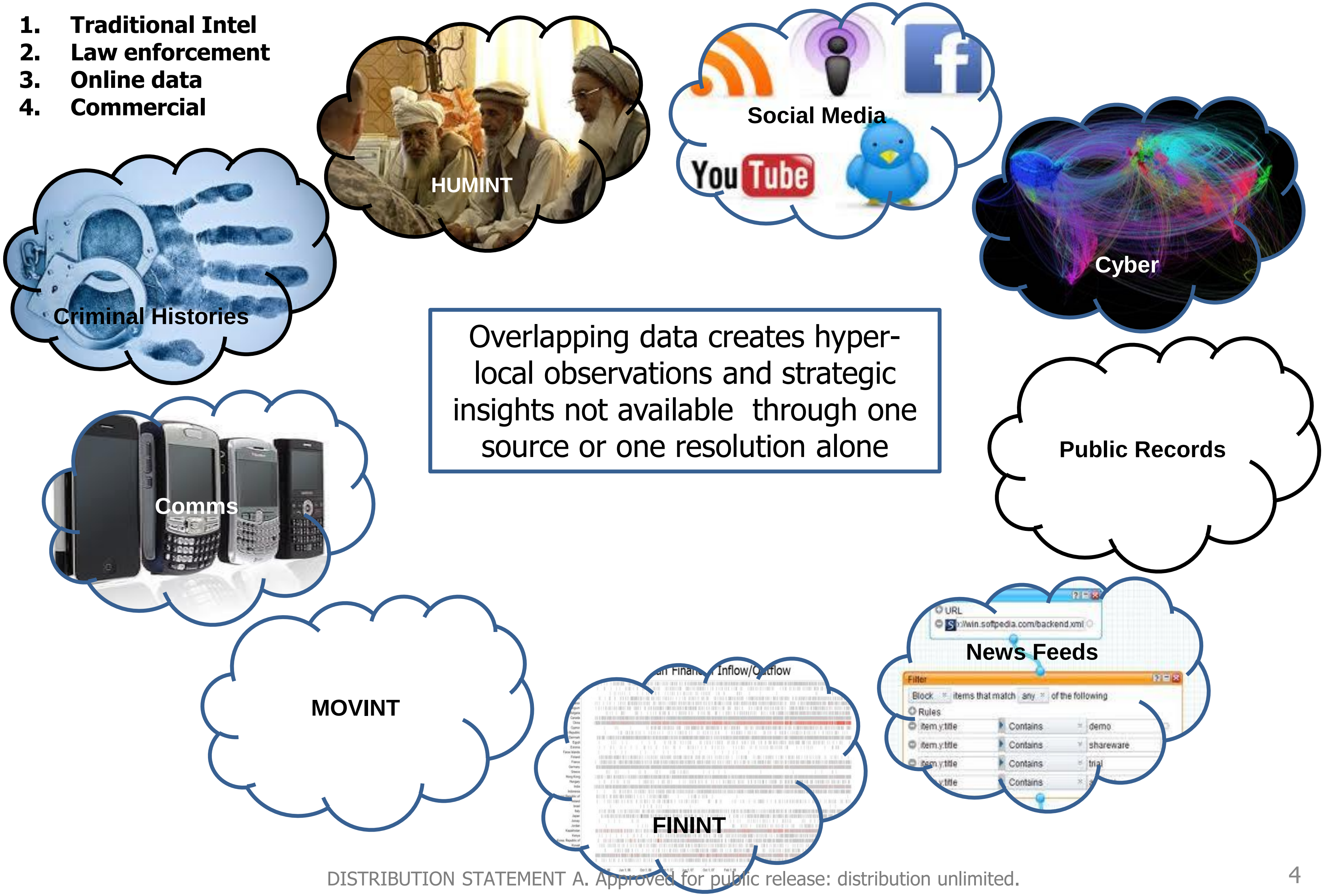
- $153B market for opiates and cocaine
- **Profits fuel terrorism, violence, corruption**

(UNODC World Drug Report 2011)



Integrated Data Sources Enhance Comprehensive Picture

1. **Traditional Intel**
2. **Law enforcement**
3. **Online data**
4. **Commercial**

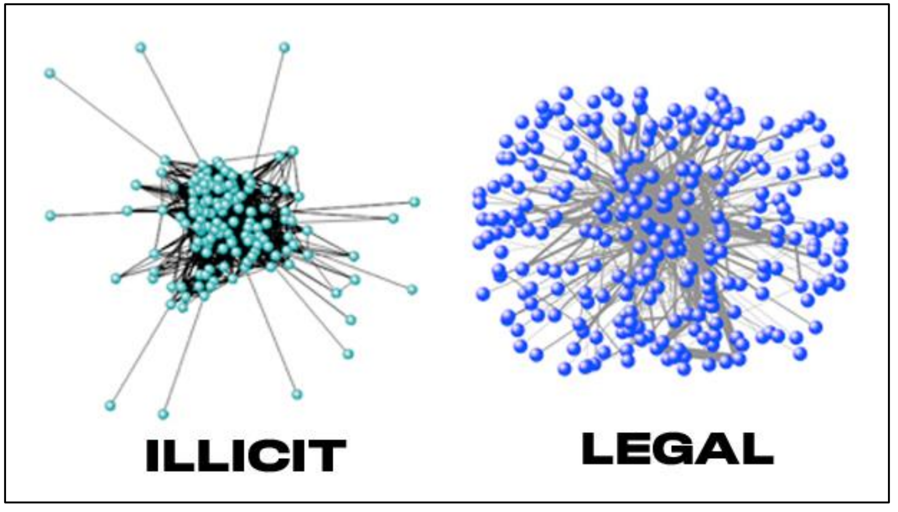


Overlapping data creates hyper-local observations and strategic insights not available through one source or one resolution alone

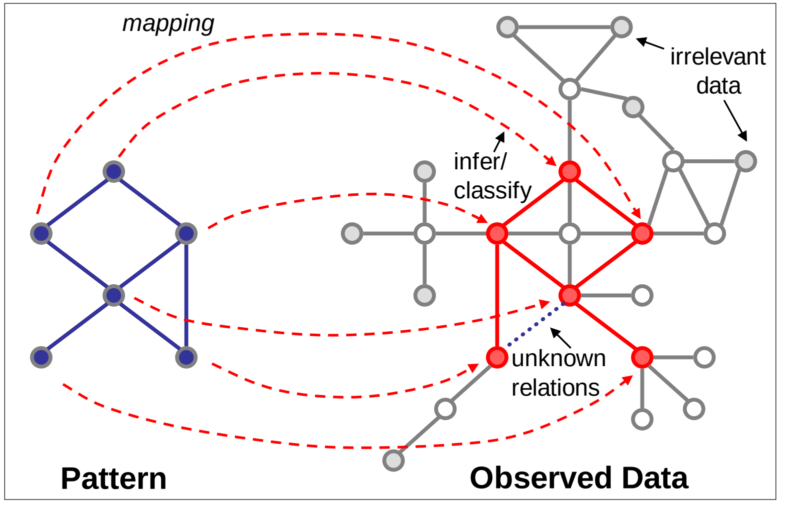
Reveal Organizations: Mismatch between expected and observed behavior

Detect exact and inexact patterns in networks to determine when predicted organizational behaviors match observations.

$$\min P(\mathbf{S}|\mathbf{D},\mathbf{M}) \cong \frac{1}{Z(\mathbf{D},\mathbf{M})} \prod_{ki} \left[p\left(a_{i,i}^{\mathbf{D}} | a_{k,k}^{\mathbf{M}}\right)\right]^{S_{ki}} \prod_{kmij} \left[p\left(a_{i,j}^{\mathbf{D}} | a_{k,m}^{\mathbf{M}}\right)\right]^{S_{ki}S_{mj}}$$



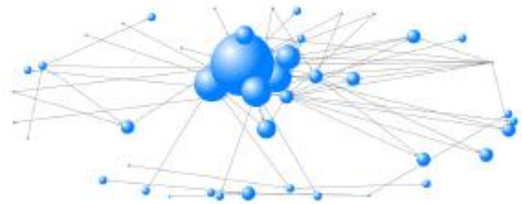
Detectable communication pattern from Enron data. A clique surrounded by leaves keeps illegal information contained within a small group.



Pattern matching detects exact and noisy matches to ultimately detect illegal communication patterns.

Enron communication networks: corruption consistently produces visible differences in individual communication

Corrupt Project 1

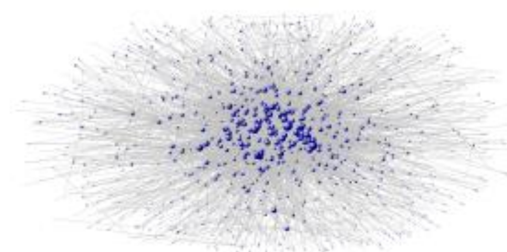


Corrupt Project 2

Corrupt Project 3

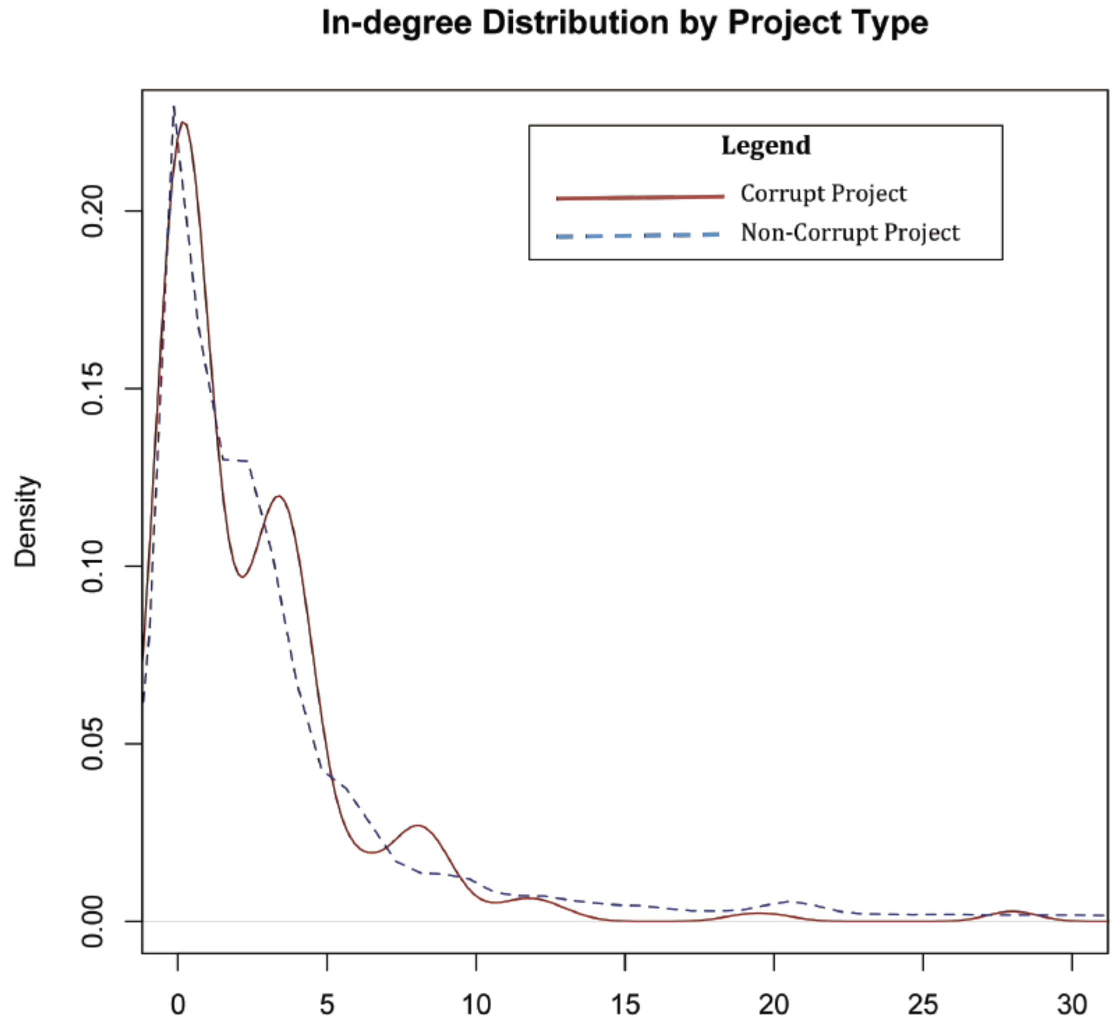
Non-Corrupt Project 1

Non-Corrupt Project 2



Non-Corrupt Project 3

Distinguishing between corrupt and noncorrupt communication requires complex network signatures



Examining a single metric such as in-degree is insufficient. Structural information appears to be required.

Signature aggregation improves probability of detection

Given

- S_x and S_y independent data sources
- $P(S_y|H) \geq P(S_y)$

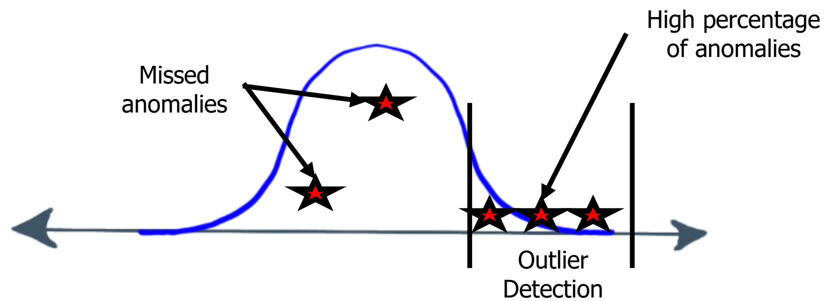
Then $P(H|S_x \cap S_y) \geq P(H|S_x)$.

Thus, the probability of detecting an event improves when you have more than one independent data source.

Example: Outlier detection for anomaly identification

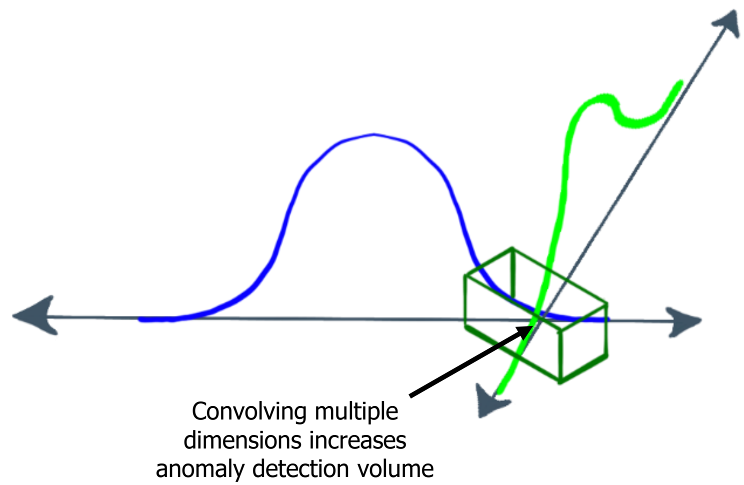
Outlier detection for identifying anomalies:

- Underlying hypothesis is that anomalies are statistical outliers along some dimension
- Commonly assume distributions are Gaussian – not necessarily true
- Low false positive rates are easily achievable AT THE EXPENSE of high false negative rates



Insight: Many anomalies are outliers along multiple dimensions; high false negative rates can be ameliorated:

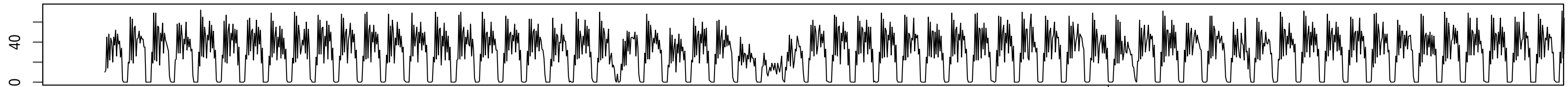
- Develop different outlier detection algorithms for multiple dimensions
- Lower the threshold for each outlier detection algorithm → increases false positives but decreases false negatives
- Convolving the different outlier detection algorithms lowers false positives without undue impact on false negatives
- Outlier detection algorithms can be combined with more sophisticated anomaly detection techniques for further enhancement



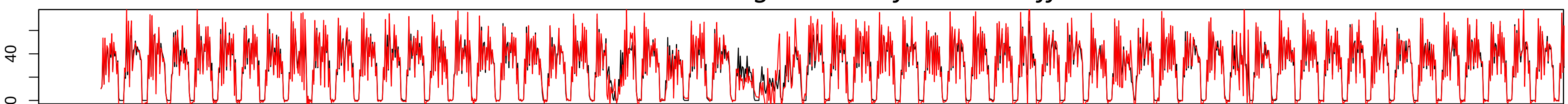
Detect Behavior: Autoregressive Integrated Moving Average (ARIMA)

- Useful for behavior detection and prediction
- Incorporates seasonal trends and patterns

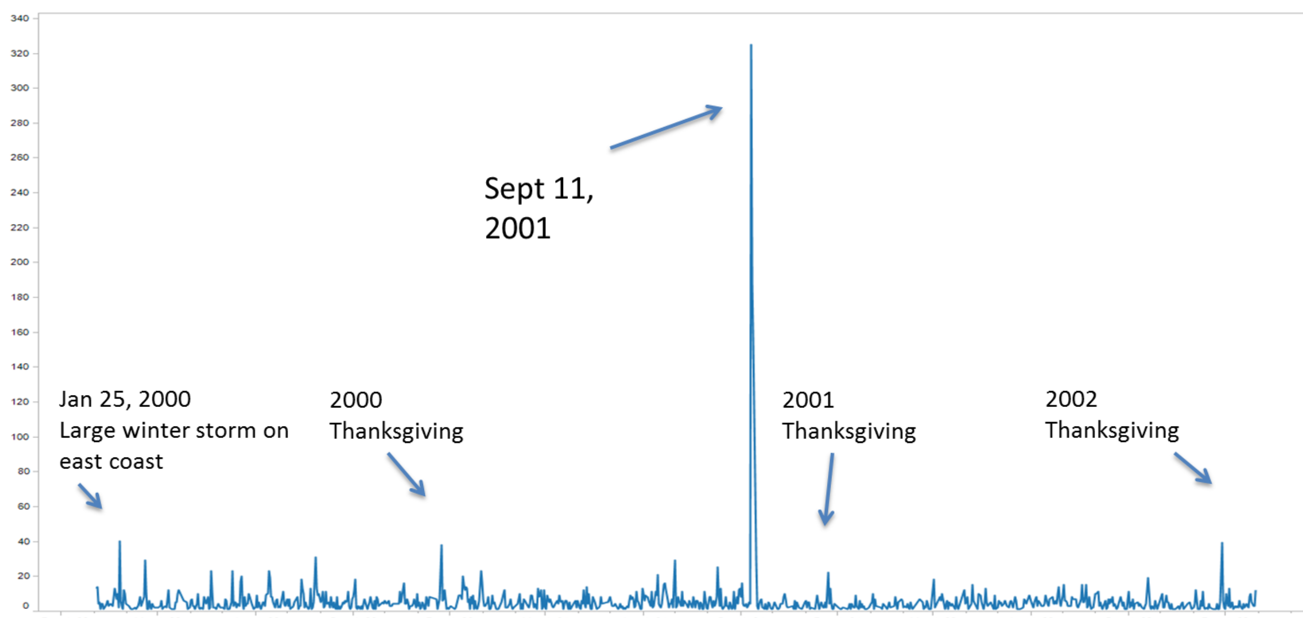
$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$



Time series showing volume of airline traffic



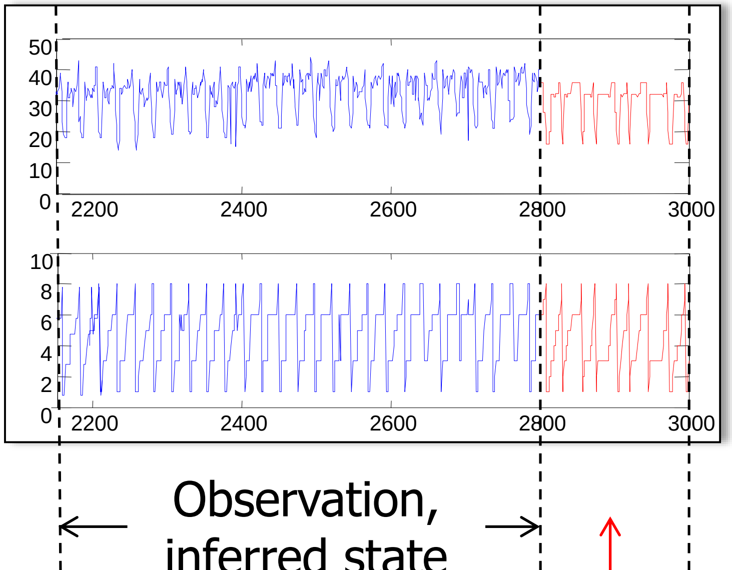
Time series showing volume of airline traffic with ARIMA overlaid in red



ARIMA Standardized Residuals over flight data from 2000-2002

Detect Behavior: Signature Detection via Dynamic Tracking

Power plant output/hour modeled as an HMM

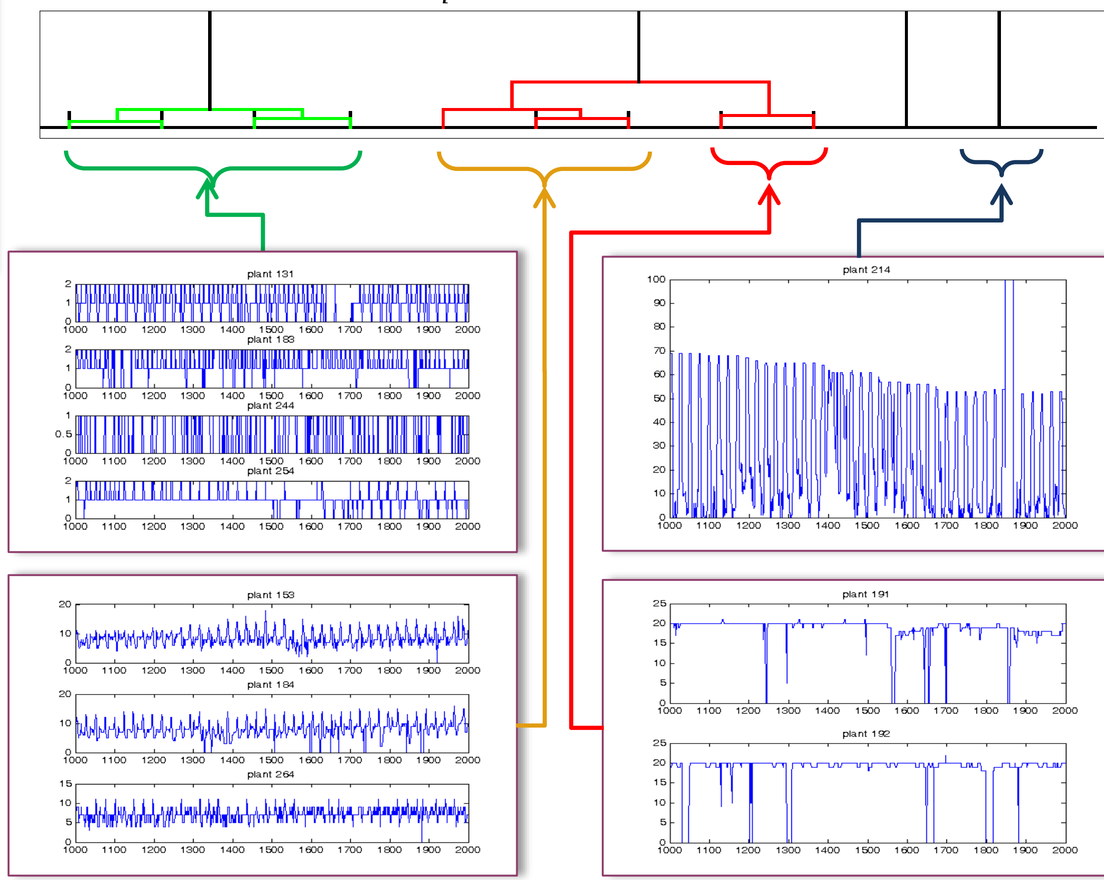


Observation, inferred state

Predicted output and HMM state

Clustering HMM models identifies classes of qualitatively similar signatures:

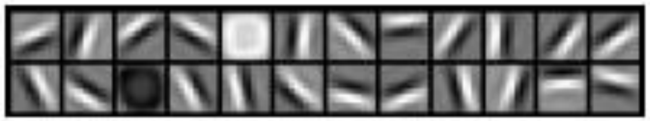
$$d_{i,j} = D(\lambda_i, \lambda_j) = \frac{1}{T_i}\left[\log\left(P(O^i \mid \lambda_i)\right) - \log\left(P(O^i \mid \lambda_j)\right)\right]$$



Detect Behavior: Unsupervised feature learning yields minimal basis for complex signals

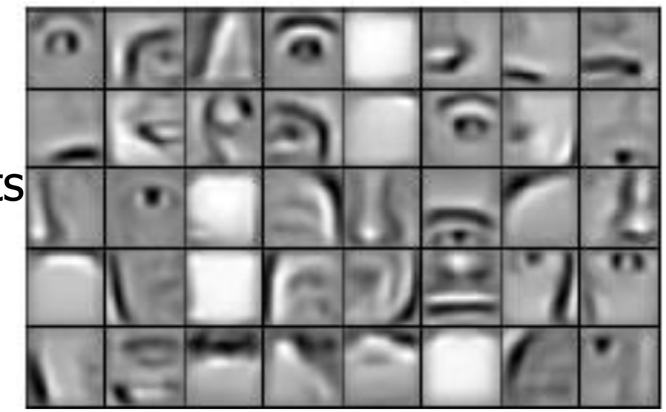
$$\min_{W_1 W_2} \sum_{i=1}^{m} \left(\left\| W_2 W_1{}^T x^{(i)} \right\|_2^2 + \lambda \sum_{j=1}^{k} \sqrt{\epsilon + H_j\left(W_1{}^T x^{(i)}\right)^2} \right)$$

Edges



→ Patterns in single data sources

Face Parts



→ Organization parts (Teams)

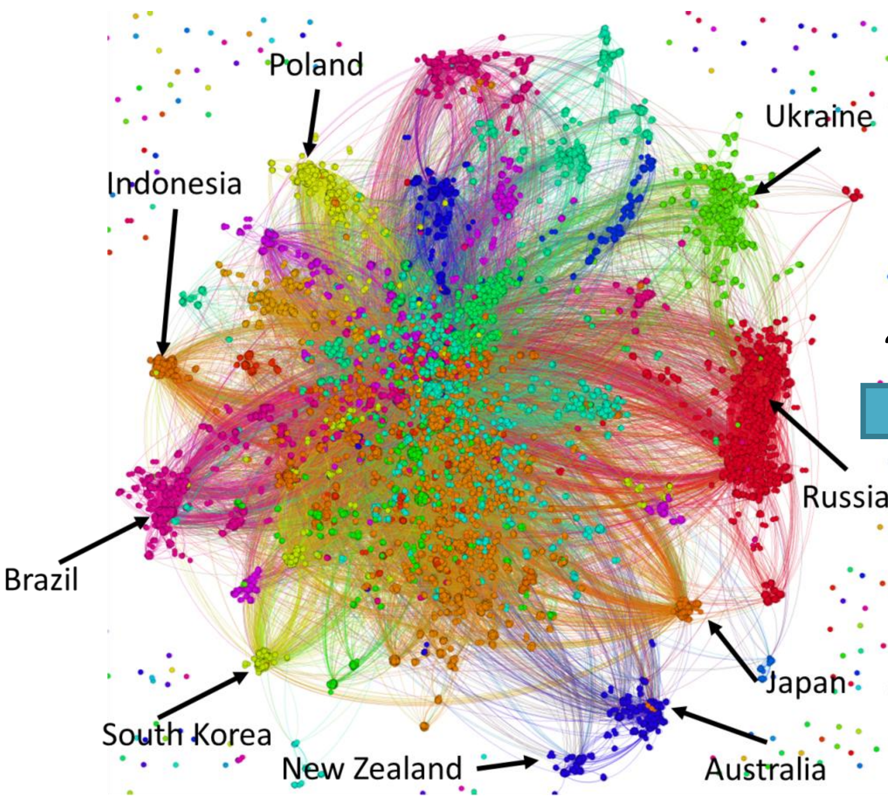
Faces

→ Organizations

Reveal Organizations: Modularity

Automated community detection based on structural characteristics

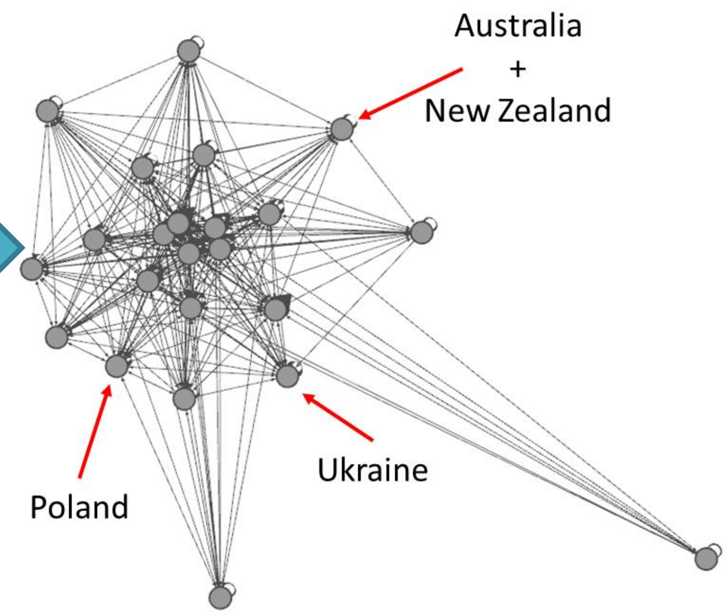
- Detects groups of nodes with high density connections amongst themselves
- Allows **hierarchical extraction of organizational** characteristics



Example modularity analytic shown on trace-route data. Each color shows a detected community, which ends up being closely aligned with a country / countries of interest.

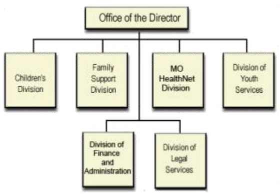
$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

Zoom Out!



Ability to summarize data at an organizational level and provide higher levels of abstraction / characterization

Generative Model Framework: Connecting Organizations to Observables



Organization Type
(Function, Location …)

↓

Structure, Internal Communications,
Resources (Inputs) & Products (Outputs)

↓

Observables, Artifacts

- New Technology
  - Constructing patterns of organizational activity via:
    - Generative models connecting organizations to observables
    - Network grammars constraining interactions between organizations
    - Transfer function views of organizations highlighting inputs and outputs
- Open Questions
  - What dimensions of an organization affect its structure, communication patterns, resources, and products?
  - How do the structure and processes affect the expected distributions of observables?
  - How do the expected observables and artifacts vary?

Modeling Organizations: Deep Generative Models Connect Organizations to Observables

Models encode expert knowledge →

Organization Type
(Function, Location …)

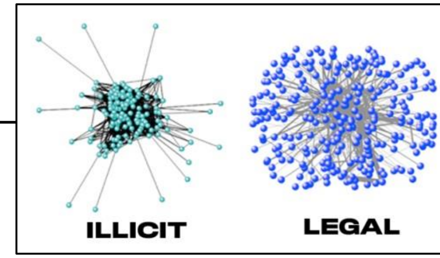
↓

Structure, Internal Communications,
Resources (Inputs) & Products (Outputs)

↓

Observables, Artifacts

Plausible observables generated via noise models ↗



$$P(Org, Structure|Obs) \propto P(Obs|Structure)P(Structure|Org)P(Org)$$

- Organizational templates from social science provide top-down information, but observed signatures differ from theoretical predictions.
- Hierarchical generative models provide mathematics for formalizing organizational theories and connecting them to realistic signatures.

Organizational Modeling System