Inferring Polyadic Events with Poisson Tensor Factorization

Aaron Schein — UMass Amherst

John Paisley — Columbia University

David M. Blei — Columbia University

Hanna M. Wallach — Microsoft Research

Motivation

- Social networks can be represented *dynamically* as a set of time-stamped interactions between actors
- Data is often limited to *dyadic interactions* (between pairs)
- Analysis are often interested in discovering dynamic structures relating multiple actors
- **Modeling goal:** Discover dynamic and typed communities of actors —i.e., *polyadic events* — from dyadic interaction data

Dyadic data in international relations

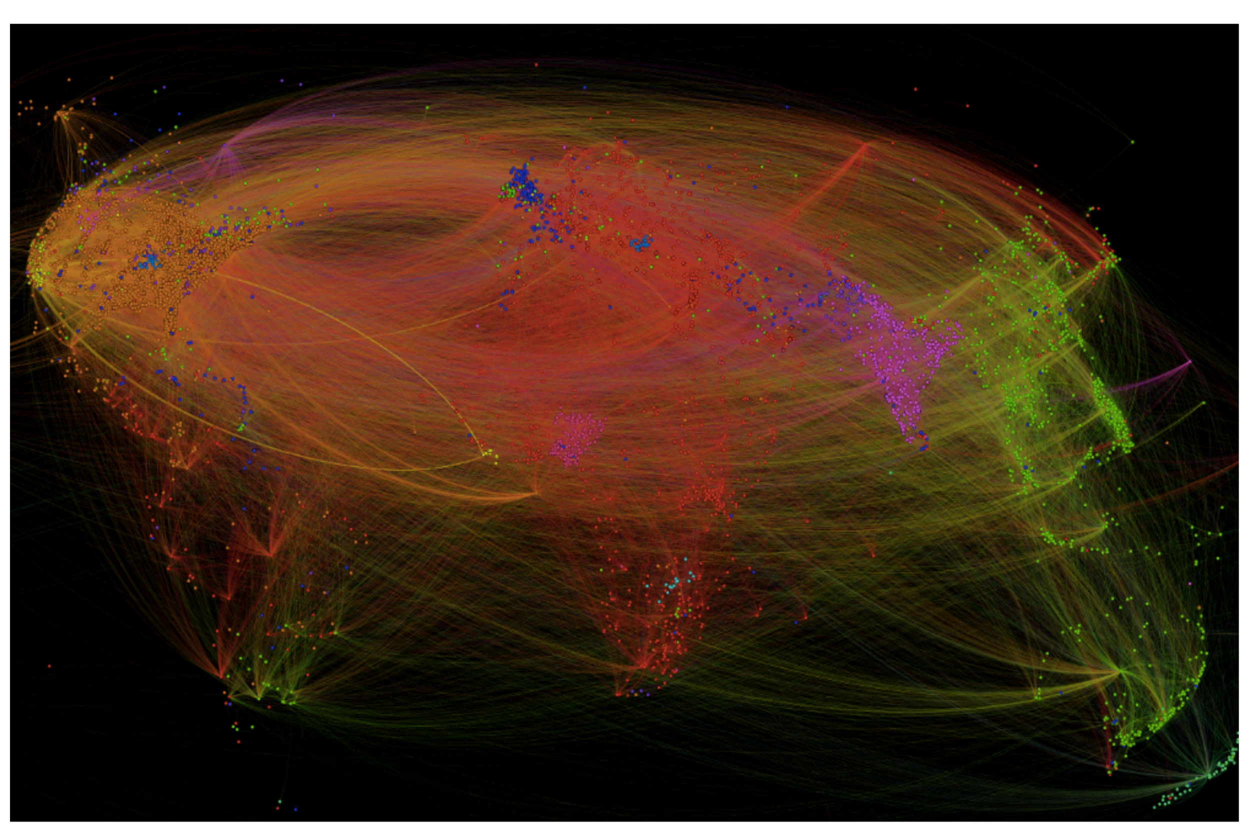
- Dyadic interaction: "**Who** did **what** to whom, when"
- Political scientists have been collecting records of country interaction to analyze patterns of international relations
- Records are automatically extracted from news articles, e.g.:

"December 8: Iranian jets bomb targets in Iraq."

12/8/14 Iran fight Iraq

Global Database of Events, Language and Tone

- GDELT is the largest database of country interaction records
- Over a ***quarter billion*** interactions from 1979 to present
- Uses the CAMEO coding scheme for actors and action-types



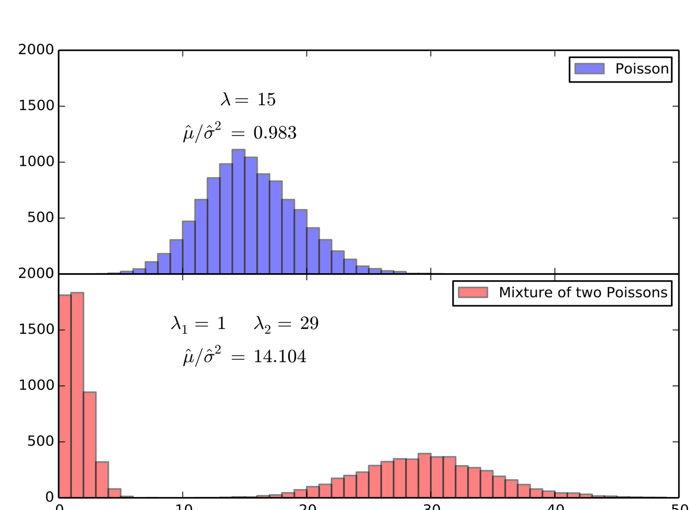
Picture © 2014 Kalev Leetaru, available on the GDELT blog.

Overdispersion in count data

- Data is *overdispersed* when the variance exceeds the mean
- The Fano factor is a measure for overdispersion:

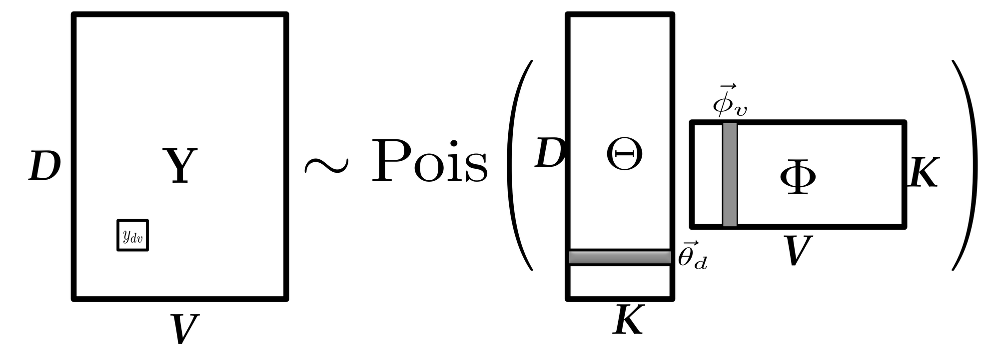
$$F = \frac{\mu}{\sigma^2}$$

- Poisson distributed counts have expected Fano factor of 1
- Overdispersion in counts can be viewed as evidence of hidden structure — i.e., a *signal-to-noise* metric
- Poisson factorization can be understood as explaining overdispersed counts via mixtures of Poissons



Poisson Matrix Factorization (PMF)

PMF is a form of nonnegative matrix factorization for count data:



where observed counts are assumed to be drawn from a Poisson centered around the dot product of two latent K-length vectors:

$$y_{dv} \sim \text{Pois}\left(\sum_{k=1}^{K} \theta_{dk}\phi_{kv}\right)$$

Poisson Tensor Factorization (PTF)

- PTF is a simple generalization of PMF to *tensors* of counts
- For dyadic interactions, the observed tensor has 4 *modes:*
  - Mode 1 — **Senders** indexed by i
  - Mode 2 — **Receivers** indexed by j
  - Mode 3 — **Action-types** indexed by a
  - Mode 4 — **Time-steps** indexed by t

$$\mathbf{Y} \equiv \{y_{ijat}\} \in \mathbb{N}^{N \times N \times A \times T}$$

- Each observed count in the tensor is assumed drawn from a Poisson centered around the multi-way product of four K-length latent vectors (one for each mode); with priors, the full *generative process* is:

$$\theta^s_{ik} \sim \text{Gamma}(a, b)$$
$$\theta^r_{jk} \sim \text{Gamma}(a, b)$$
$$\psi_{ak} \sim \text{Gamma}(c, d)$$
$$\delta_{tk} \sim \text{Gamma}(e, f)$$
$$y_{ijat} \sim \text{Pois}\left(\sum_{k=1}^{K} \theta^s_{ik}\theta^r_{jk}\psi_{ak}\delta_{tk}\right)$$

Mean field variational inference

Our variational inference algorithm allows us to ***efficiently*** fit the model to ***very large*** datasets. The goal of inference is to compute the posterior distribution:

$$P(\mathbf{\Theta^s}, \mathbf{\Theta^r}, \mathbf{\Psi}, \mathbf{\Delta} \mid \mathbf{Y})$$

In variational, we *approximate* the posterior by *optimizing* a tight lower bound on the true joint probability:

$$\mathcal{B} = \mathbb{E}_q[\ln P(\mathbf{Y}, \mathbf{\Theta^s}, \mathbf{\Theta^r}, \mathbf{\Psi}, \mathbf{\Delta})] + H(q)$$

where q is an *instrumental distribution* over the latent factors that fully factorizes over all latent variables. We exploit the Poisson-Multinomial relationship to form an instrumental distribution that is conditionally conjugate:

$$q(\vec{z}_{ijat}) = \text{Multinomial}(y_{ijat}, \vec{\phi}_{ijat})$$
$$q(\theta^s_{ik}) = \text{Gamma}(\alpha^s_{ik}, \beta^s_{ik})$$
$$q(\theta^r_{jk}) = \text{Gamma}(\alpha^r_{jk}, \beta^r_{jk})$$
$$q(\psi_{ak}) = \text{Gamma}(\gamma_{ak}, \chi_{ak})$$
$$q(\delta_{tk}) = \text{Gamma}(\rho_{tk}, \nu_{tk})$$

Optimizing the lower bound can then performed simply and efficiently via coordinate ascent on the variational parameters.

Guided exploration

- We fit to GDELT data from 2012 with weekly binning, top-level action-types, and all countries
  T = 52 A = 20 N = 219
- Exploration of the results is guided by overdispersion as a signal-to-noise metric
- **Example 1:** The most overdispersed *receivers* are:
  1. *USA* 2. *Israel* 3. *Palestine* 4. *Myanmar* 5. *Equador*
  —> Find which components a receiver is most active in (Figure 1)
  —> Explore and interpret those components (Figures 3-4)
- **Example 2:** The most over dispersed *action-types* are:
  1. *Consult* 2. *Make Statement* 3. *Fight*
  —> Find which components an action-type is most active in (Figure 2)
  —> Explore and interpret those components (Figures 5-8)

Figure 1: Factors $\vec{\theta^r}_{j\cdot}$ for receiver *Myanmar*

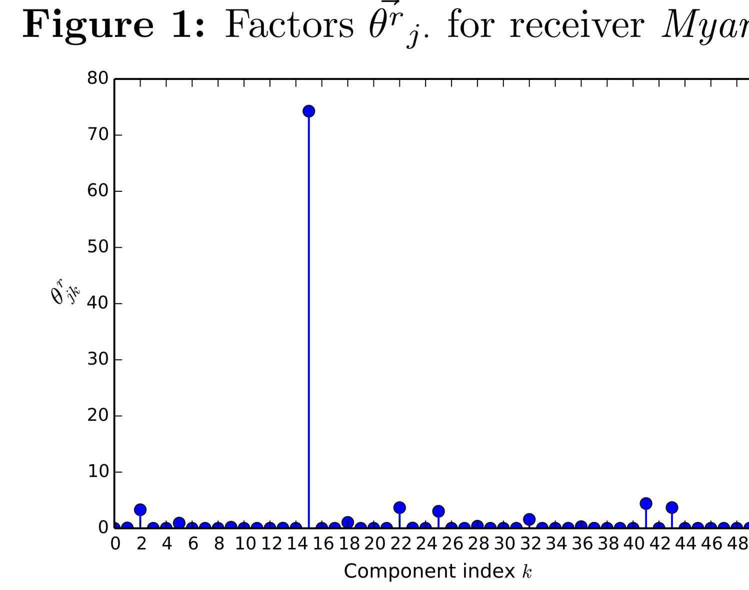


Figure 2: Factors $\vec{\psi}_{a\cdot}$ for action-type *Fight*

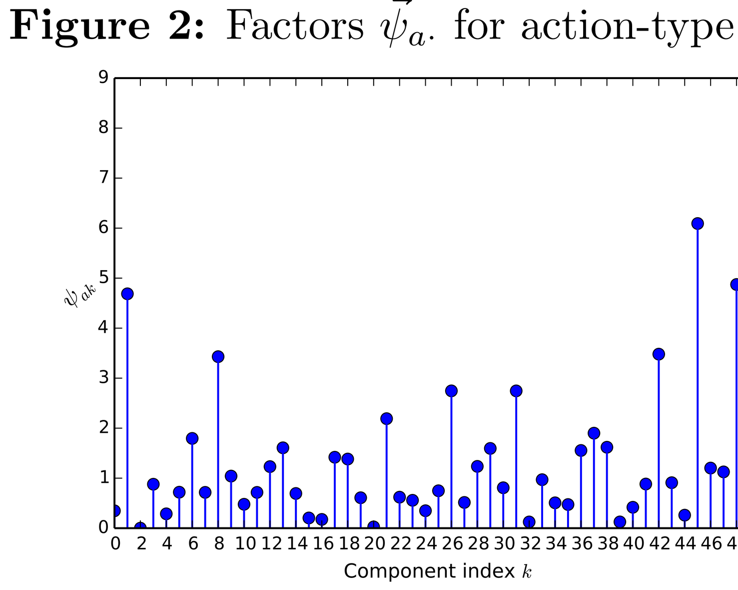


Figure 3: Julian Assange seeks asylum in Ecuador

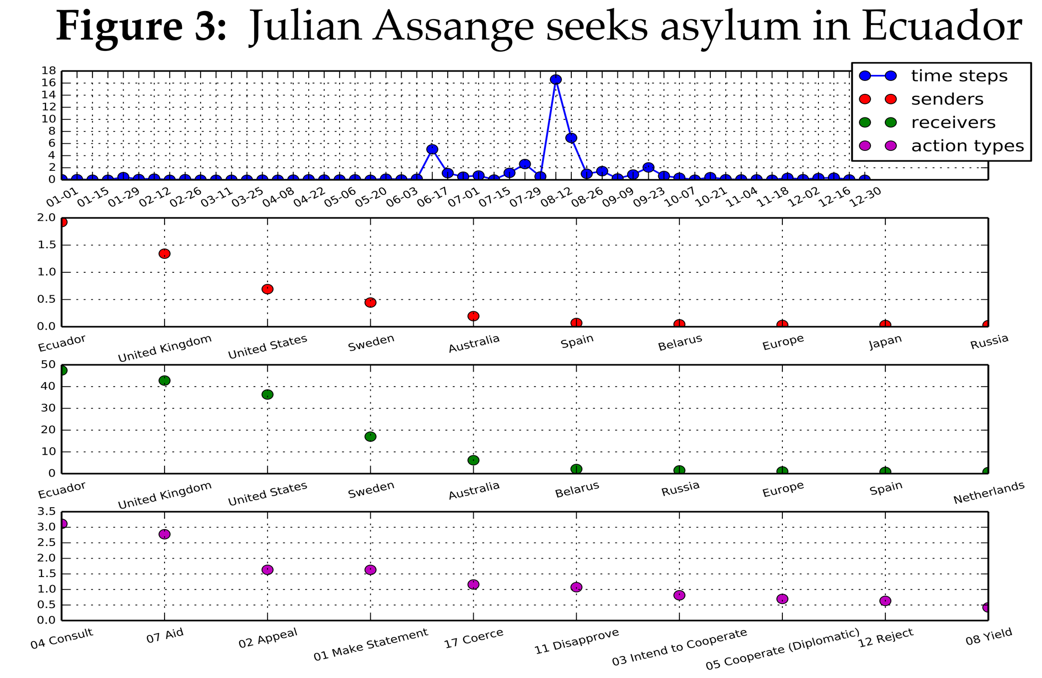


Figure 4: Obama administration's ``Pivot to Asia''

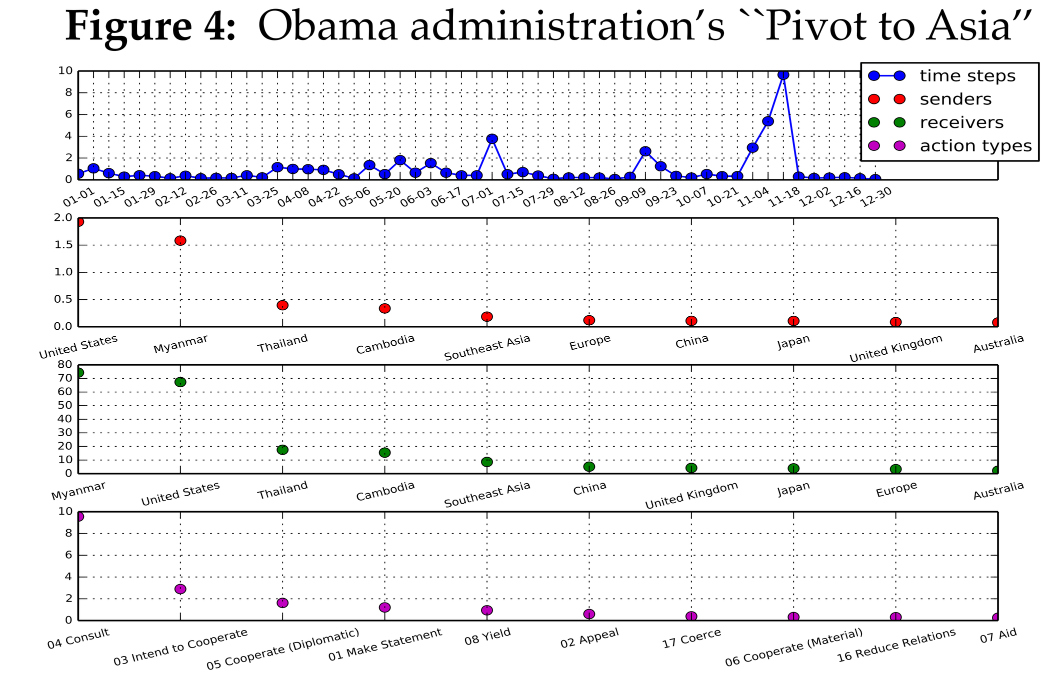


Figure 5: Benghazi attack on US embassy in Libya

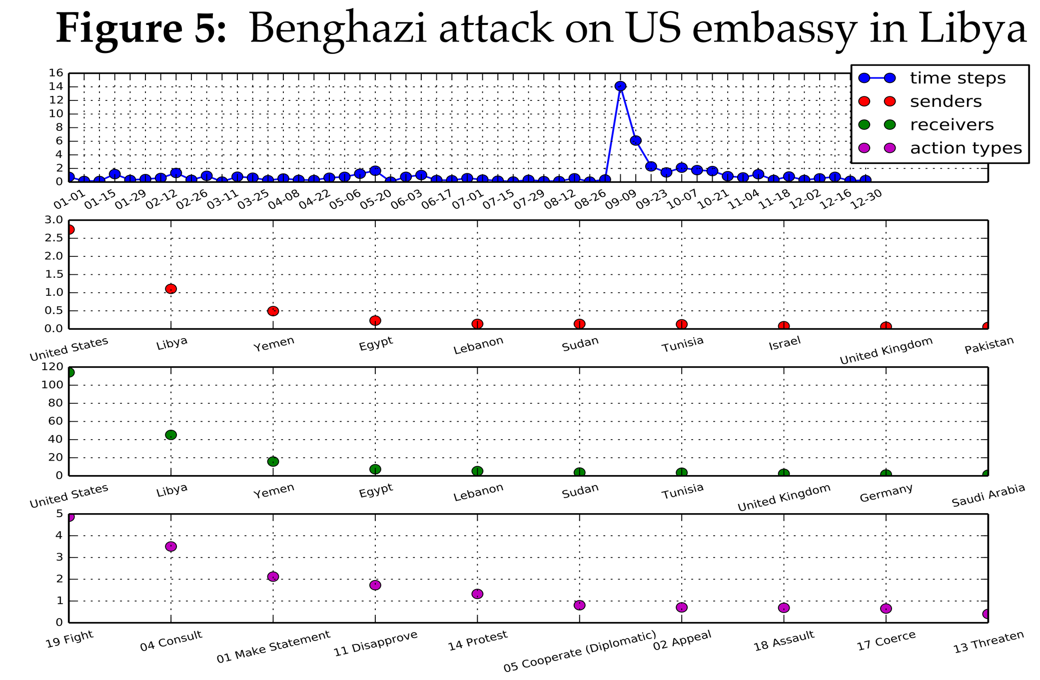


Figure 6: Syrian civil war

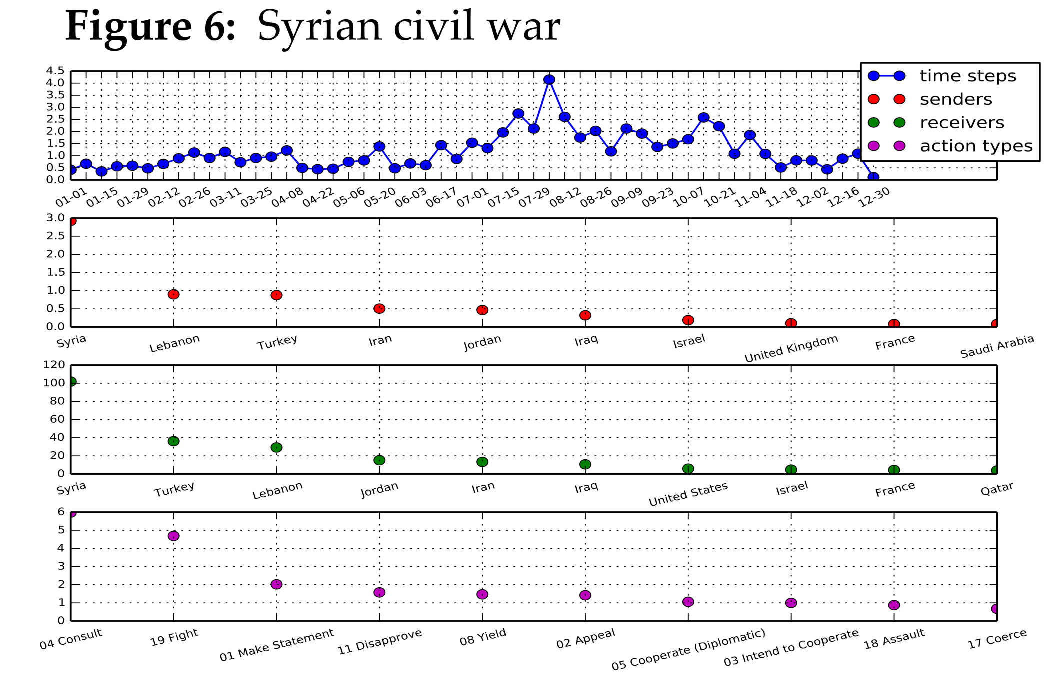


Figure 7: Operation Pillar of Defense

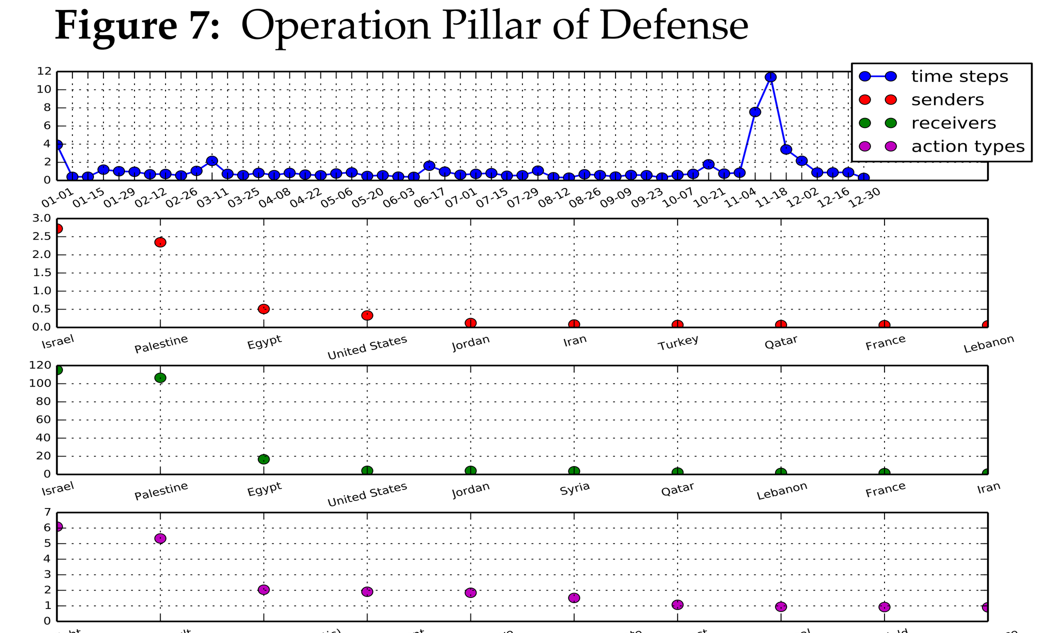


Figure 8: War in Afghanistan-Pakistan