Dynamic Models for Graphs

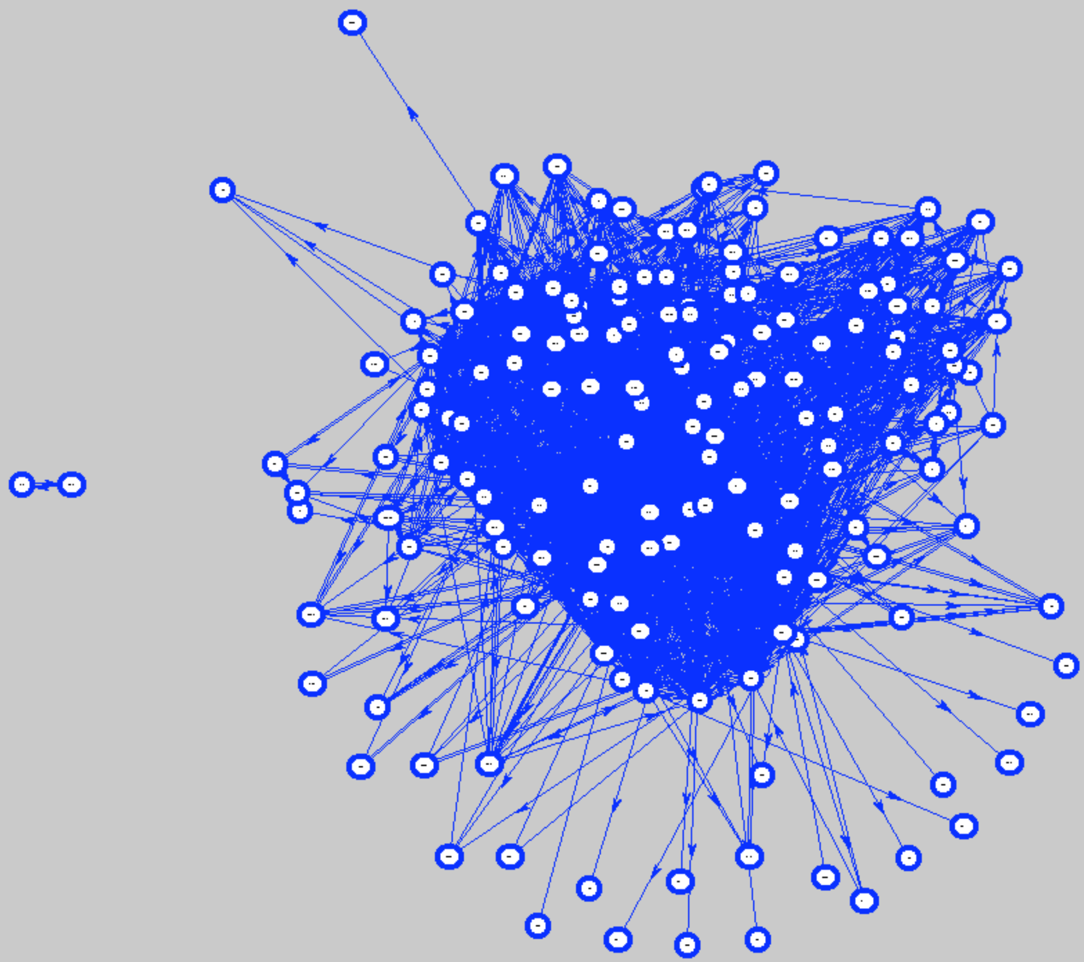
William F. Szewczyk

Math Research Group
National Security Agency
wfszewc@afterlife.ncsc.mil

The Data

- Central Asia (CASIA) database from the Kansas Event Data Survey (http://www.ku.edu/keds/data.html)
- Events as reported in Reuters newswire
- 139 state and non-state actors
- Events from May 1989 through July 1999

What's Wrong with This Picture?



The problem with most probabilistic models for graphs is that they don't account for any graph dynamics.

The p^*-model

Let $G = \{V, E\}$ be a directed graph with a vertex set $V = \{v_1, \ldots, v_g\}$, and edge set $E = \{e_{ij}\}$ where $e_{ij} = 1$ if vertex v_i sends a link to vertex v_j and 0 otherwise for $i = 1, \ldots, g, j = 1, \ldots, g, i \neq j$. Model the logit of $P(e_{ij} = 1)$ as

$$\text{logit}(e_{ij}) = \log\left(\frac{P(e_{ij} = 1)}{P(e_{ij} = 0)}\right) = \alpha_i + \beta_j + \gamma.$$

Some Attempts at Dynamism

There appear to be two main ways to introduce some dynamics into a graph:

- Assume that each graph is an independent sample from some (possibly unknown) distribution and look for changes.
- Use an exponential smoothing scheme to weight recent activity more than activity in the past.

What's really needed is a "Kalman Filter" for graphs.

Review of Linear Models

Let $\mathbf{Y} = (y_1, \ldots, y_t)'$ be a series of observations. Suppose there's a set of unknown parameters, θ, such that for a known design matrix, $\mathbf{F}$,

$$\mathbf{Y} = \mathbf{F}'\theta + \mathbf{V},$$

where $\mathbf{V} = (\nu_1, \ldots, \nu_t)'$ is a vector of iid disturbance terms.

If one assumes that $\nu_i \sim N(0, \sigma^2)$, then one has at their disposal all the standard regression tools.

Dynamic Linear Models

$$y_t = \mathbf{F}'_t \theta_t + \nu_t$$
$$\theta_t = \mathbf{G}_t \theta_{t-1} + \omega_t$$

where ω_t is a disturbance term uncorrelated with θ_{t-1} and ν_t.

Now how does one estimate θ_t?

The Basic Idea

1. Take what we know now (D_t)
2. Predict what we will see next
3. See what we see next
4. See how far off we are
5. Fix our mistakes
6. Iterate

An Example

$$
\begin{aligned}
y_t &= \mathbf{F}_t'\theta_t + \nu_t & \nu_t &\sim N(0, V) \\
\theta_t &= \mathbf{G}_t\theta_{t-1} + \omega_t & \omega_t &\sim N(0, \mathbf{W}_t) \\
(\theta_{t-1}|D_{t-1}) &\sim N(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}) & & \\
(\theta_t|D_{t-1}) &\sim N(\mathbf{a}_t, \mathbf{R}_t) & \mathbf{a}_t &= \mathbf{G}_t\mathbf{m}_{t-1} \\
& & \mathbf{R}_t &= \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t \\
(Y_t|D_{t-1}) &\sim N(f_t, Q_t) & f_t &= \mathbf{F}_t'\mathbf{a}_t \\
& & Q_t &= \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + V \\
e_t &= y_t - f_t & \mathbf{A}_t &= \mathbf{R}_t\mathbf{F}_t/Q_t \\
\mathbf{m}_t &= \mathbf{a}_t + \mathbf{A}_t e_t & \mathbf{C}_t &= \mathbf{R}_t - \mathbf{A}_t\mathbf{A}_t'Q_t
\end{aligned}
$$

The Linear Model Revisited

Recall the linear model $y = \mathbf{F}'\theta + \nu$ with $\nu \sim N(0, \sigma^2)$ One could decompose this into three components

1. A RANDOM COMPONENT: $Y \sim N(\mu, \sigma^2)$, where, $\mu = E(Y)$.
2. A SYSTEMATIC COMPONENT: A linear predictor $\eta = \mathbf{F}'\theta$.
3. A LINK COMPONENT: $g(\mu) = \eta$, in this case the identity.

So What?

1. RANDOM COMPONENT: Let $Y \sim \text{Bin}(n, \mu)$, where the probability of success is μ.
2. SYSTEMATIC COMPONENT: $\eta = \mathbf{F}'\theta$.
3. LINK COMPONENT: $g(\mu) = \log \frac{\mu}{1-\mu}$

Now we talking about *Generalized Linear Models*.

Exponential Family of Distributions

If the density of y can be written in the form

$$f_Y(y;\theta,\phi) = \exp\left((y\theta - b(\theta))/a(\phi) + c(y,\phi)\right),$$

for specific functions $a(.), b(.)$, and $c(.)$, then it is said to be of the exponential family.

Fitting GLMs

Fitting GLMs is accomplished by using an iteratively reweighted least squares algorithm. Let $\hat{\eta}_0$ be the current estimate of the linear predictor, and $\hat{\mu}_0$ the corresponding fitted response value. Form the adjusted dependent value

$$z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\left(\frac{d\eta}{d\mu}\right)_0.$$

Do a weighted regression of z_0 onto $\mathbf{F}$ with quadratic weights

$$W_0^{-1} = \left(\frac{d\eta}{d\mu}\right)_0^2 V_0$$

to obtain new estimates of θ and η.

Dynamic GLM

$$y_t = g^{-1}(\mathbf{F}_t'\theta_t)$$

$$\theta_t = \mathbf{G}_t\theta_{t-1} + \omega_t$$

The Basic Idea

1. Take what we know now (D_t)
2. Predict what we will see next
3. See what we see next
4. See how far off we are
5. Fix our mistakes
6. Iterate

Going Forward

$$y_t = g^{-1}(\mathbf{F}_t'\theta_t)$$

$$\theta_t = \mathbf{G}_t\theta_{t-1} + \omega_t \qquad \omega_t \sim N(0, \mathbf{W}_t)$$

$$(\theta_{t-1}|D_{t-1}) \sim N(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$$

$$(\theta_t|D_{t-1}) \sim N(\mathbf{a}_t, \mathbf{R}_t) \qquad \mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1}$$

$$\mathbf{R}_t = \mathbf{G}_t\mathbf{C}_{t-1}\mathbf{G}_t' + \mathbf{W}_t$$

$$\eta_{t|t-1} = \mathbf{F}'\mathbf{a}_t$$

Going Backwards

As with GLMs, DGLMs use Fisher scoring to update the parameters. Let

$$\mathbf{u}_t(\theta_t) = \partial l_t(\theta_t)/\partial\theta_t = \mathbf{F}'\mathbf{H}_t(\theta_t)\mathbf{\Sigma}_t^{-1}(\theta_t)(y_t - \mu_t(\theta_t))$$

and

$$\mathbf{U}_t(\theta_t) = E(-\partial^2 l_t(\theta_t)/\partial\theta_t\partial\theta_t'|\theta, D_{t-1}) = \mathbf{F}'\mathbf{H}_t(\theta_t)\mathbf{\Sigma}_t^{-1}(\theta_t)\mathbf{H}_t'(\theta_t)\mathbf{F},$$

where $\mu_t(\theta_t) = g^{-1}(\eta_t)$ is the conditional expectation, $\mathbf{\Sigma}_t^{-1}(\theta_t)$, the conditional covariance matrix, and $\mathbf{H}_t(\theta_t)$, the Jacobian, then

$$\mathbf{C}_t = (\mathbf{C}_{t-1}^{-1} + \mathbf{U}_t(\theta_t))^{-1} \qquad \theta_t = \mathbf{a}_t + \mathbf{C}_t\mathbf{u}_t.$$

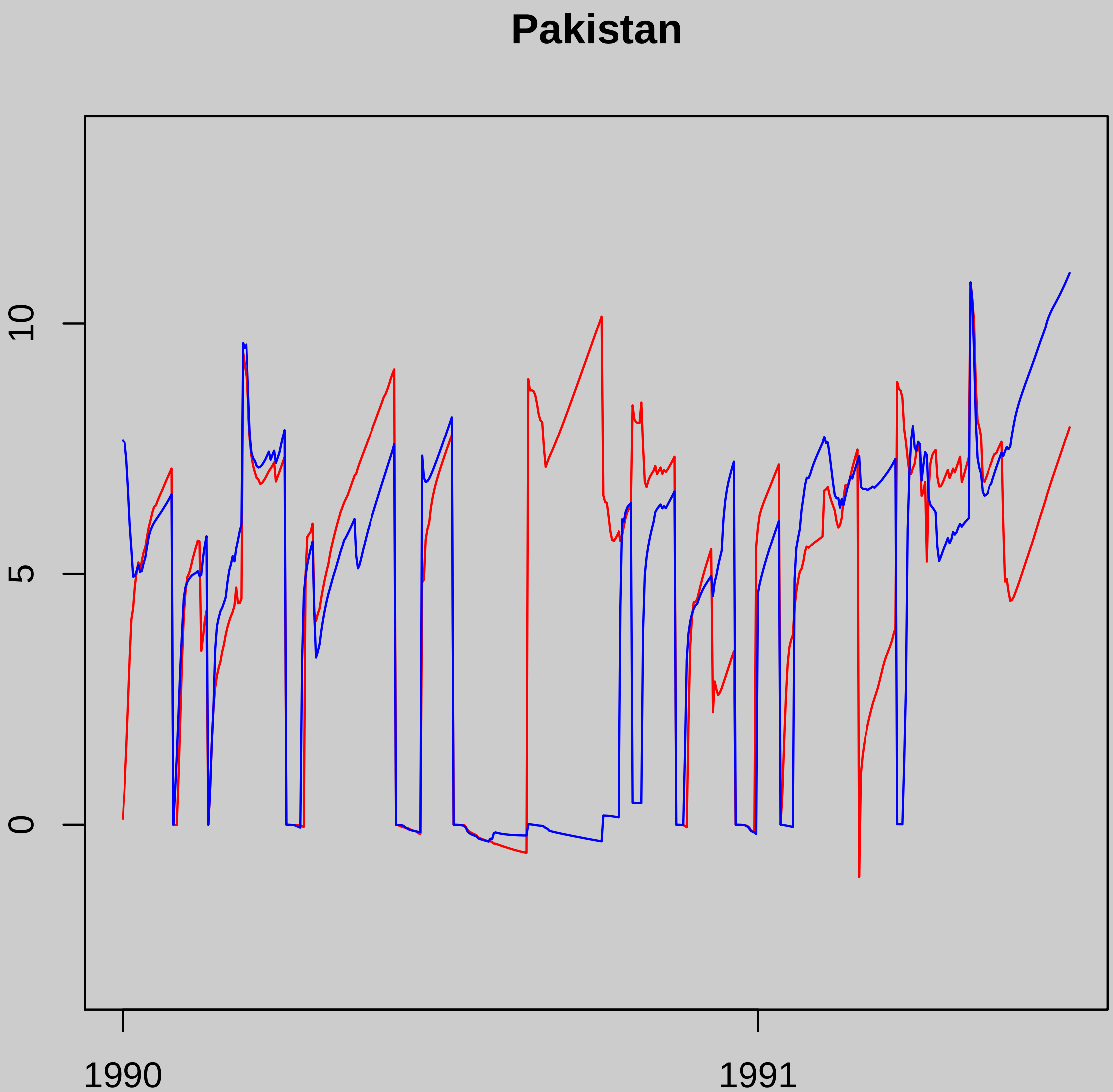
Remember This Guy?

Let $G = \{V, E\}$ be a directed graph with a vertex set $V = \{v_1, \ldots, v_g\}$, and edge set $E = \{e_{ij}\}$ where $e_{ij} = 1$ if vertex v_i sends a link to vertex v_j and 0 otherwise for $i = 1, \ldots, g, j = 1, \ldots, g, i \neq j$. Model the logit of $P(e_{ij} = 1)$ as

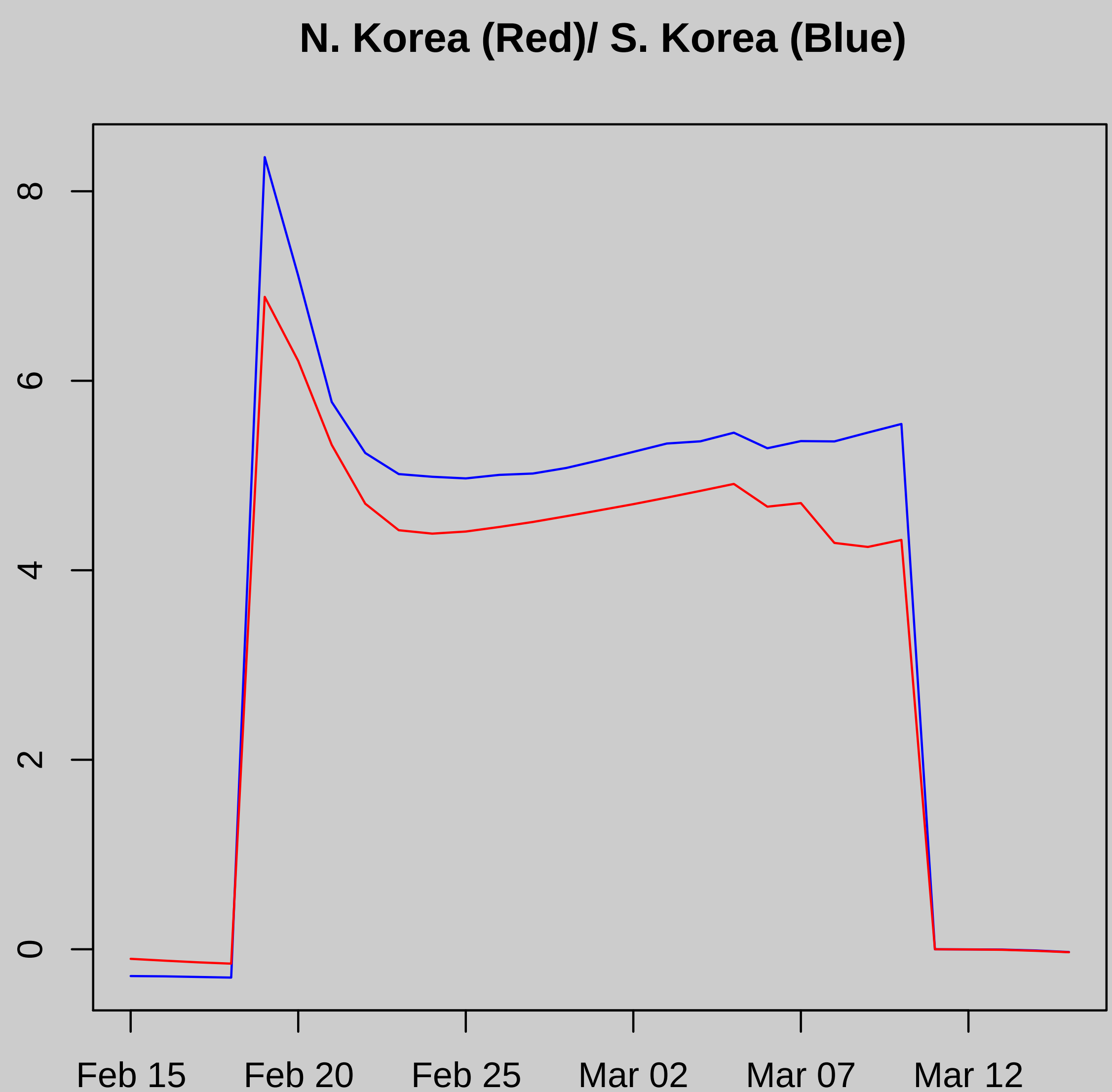
$$\text{logit}(e_{ij}) = \log\left(\frac{P(e_{ij} = 1)}{P(e_{ij} = 0)}\right) = \alpha_i + \beta_j + \gamma.$$

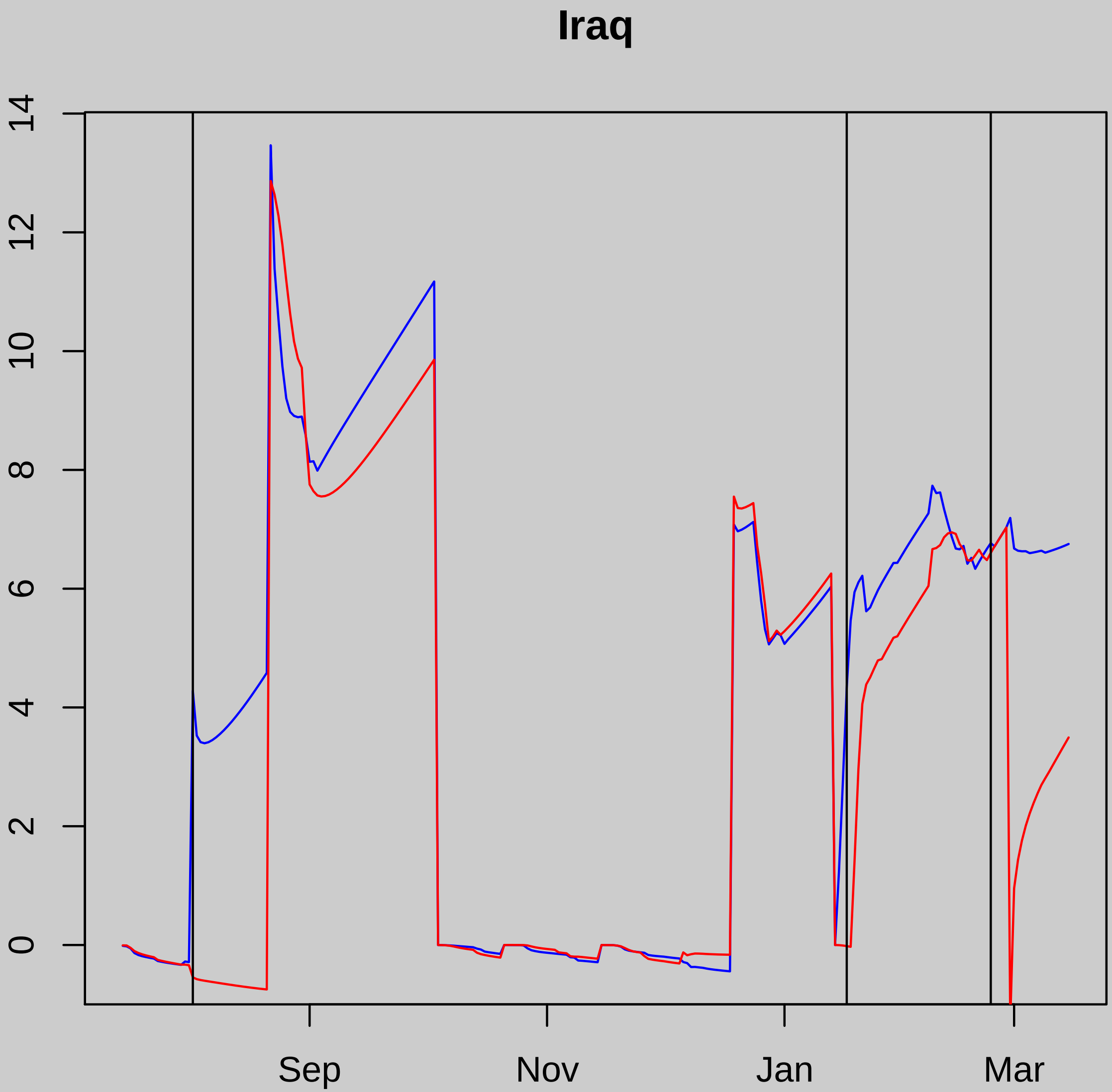
A Simple Random Walk Model

- $\mathbf{F}_t = \mathbf{F}$, a sparse $(2g - 1) \times (g^2 - g)$ matrix
- $\mathbf{G}_t = \mathbf{G}$, the $(2g - 1) \times (2g - 1)$ identity matrix
- Precision decreases by 20% at each time step
- Initialize the algorithm by fitting a p^*-model to the first graph
- Move forward in time using a DGLM with binomial errors
- If a goodness-of-fit model indicates a poor fit after some time, reinitialize

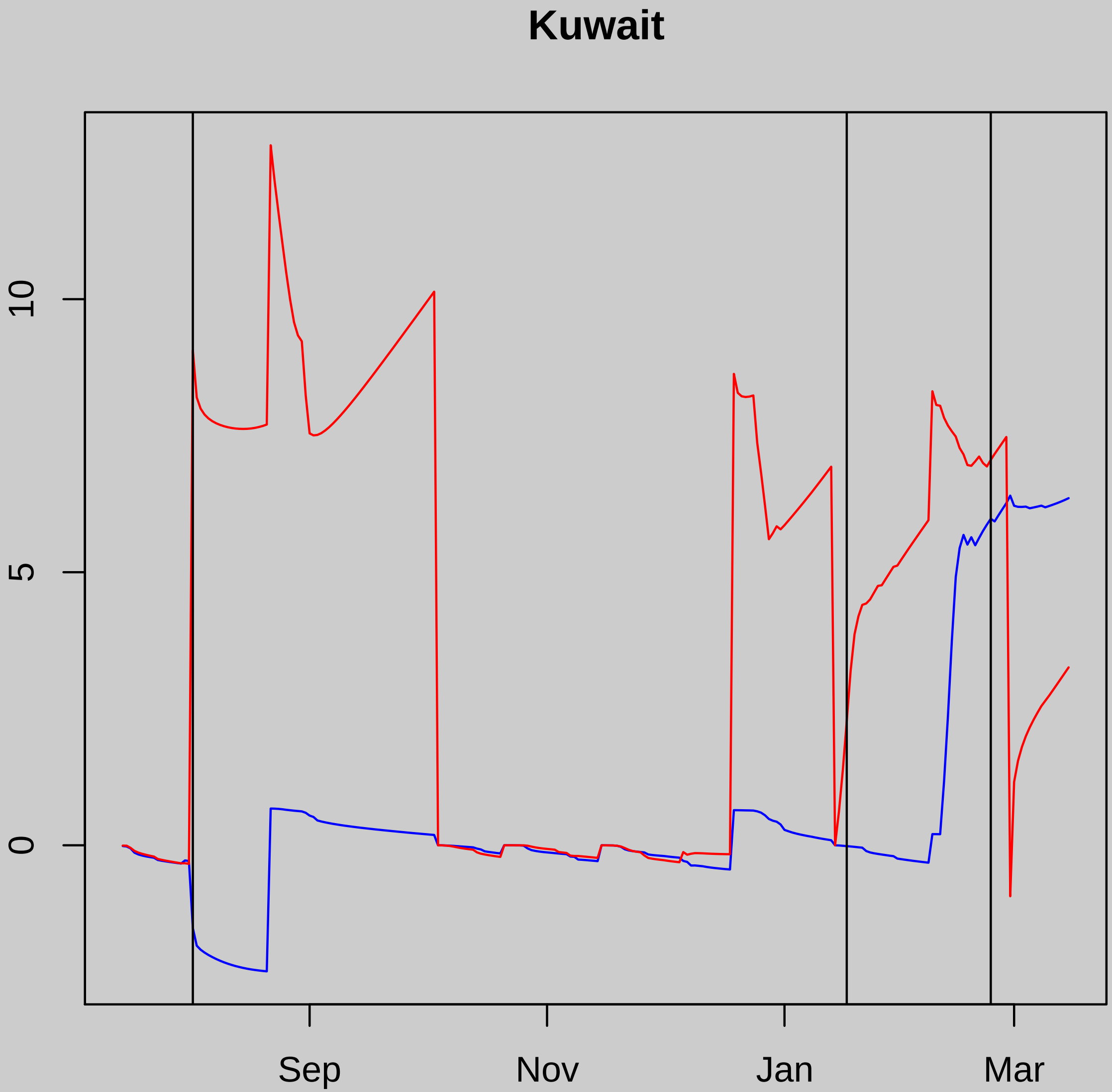


Pakistan
10
5
0
1990
1991


N. Korea (Red)/ S. Korea (Blue)
0
2
4
6
8
Feb 15
Feb 20
Feb 25
Mar 02
Mar 07
Mar 12



Iraq
14
12
10
8
6
4
2
0
Sep
Nov
Jan
Mar


Kuwait
10
5
0
Sep
Nov
Jan
Mar

References

Anderson, C. J., S. Wasserman, and B. Crouch (1999). A p^* Primer: Logit Models for Social Networks. *Social Networks 21*, 37–66.

Cortes, C., D. Pregibon, and C. Volinsky (2004). Computational Methods for Dynamic Graphs. *Journal of Computational and Graphical Statistics 12*, 950–970.

Fharmeir, L. (1992). Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models. *Journal of the American Statistical Association 87*, 501–509.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (Second ed.). Boca Raton: Chapman & Hall/CRC.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models* (Second ed.). New York: Springer-Verlag.