Trajectories of Involvement Trajectories on a Social Media Platform

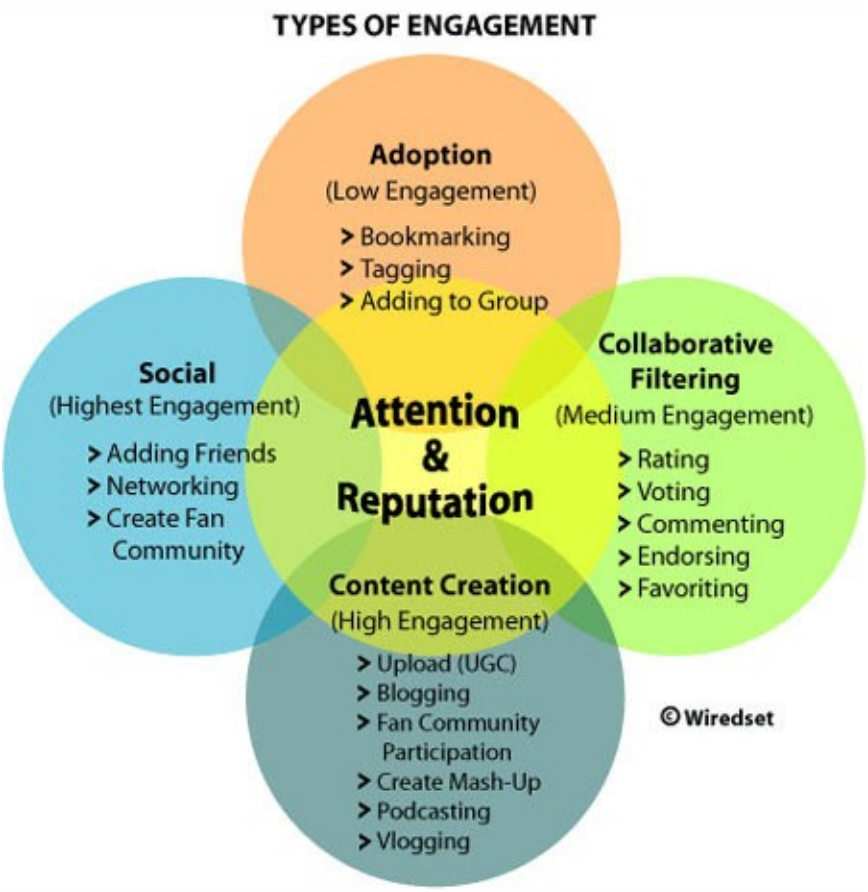
Lynd Bacon, Loma Buena Associates
Danielle Murray, Shutterfly Inc.
Peter Lenk, University of Michigan

Joint Statistical Meetings
Vancouver B.C. Canada
July 31- August 5, 2010

What is "*User* Engagement?"

- A fuzzy (and latent) construct, usually inferred from behavioral data
- A global design goal in many applications
- Various proposed metrics, from simple to complex:
  - Comscore (2007) : number of unique visits
  - Nielsen (2007) : total number of minutes
  - Forrester Research (2008) : multiple indicators, including time spent, page views, navigation

Multiple Facets of Social Media Engagement



Anthony Mayfield circa 2006. See http://www.johnniemoore.com/blog/archives/2006_11.php

Project Objectives

- Data are from Shutterfly, Inc. (“SFLY”)
- Model user involvement over time as a latent variable reflected in multiple fallible indicators
- Shed some light on how involvement develops, how it can be encouraged
- Preferably,
  - Unidimensional representation of involvement, or at least low-dimensional
  - Scalable to millions of users
  - Computation in real time not necessary
  - Predictive of sales

Some Prior Work

- Netzer, Lattin, Srinivasan
  - Hidden Markov model
  - Alumni contributions application
  - Marketing Science 2008
- Bacon, Murray & Lenk
  - IRT-type latent trait model with covariates
  - Data from Shutterfly
  - Data modeled were transformations of aggregates over time
  - Joint Statistical Meetings Proceedings, 2009

The Data at Hand

- 2,000 users registering on site during week one week
- Disguised data
  - Add & subtract Poisson RVs to activity times
- Weekly aggregates of 51 different behaviors, e.g. uploads, project creations, purchase quantities and amounts, etc.
  - Aggregate behaviors into “Upload,” “Create,” and “Purchase”
- Scant amount of user data and marketing mix
- Define cohort of 509 subjects with 5 or more activities

Features of Current Model

- Multiple, manifest indicators of involvement
- Dynamic, segmentation model for latent involvement
  - Individual-specific involvement measures
  - Involvement evolves over time
  - Combines dynamic linear model with hidden Markov process
  - Regimes correspond to increasing levels of involvement
- Covariates in both the observational and state models
  - Current data lack marketing mix

Model Part 1: Observable Equation

- User i and time t

$$y_{it} = \theta_{it}\alpha + x_{1it}\beta_i + \varepsilon_{it} \text{ and } \varepsilon_{it} \sim N_m(0, \Sigma)$$

- y_{it} is vector of manifest variables
- θ_{it} is user i's latent involvement at time t
- α is vector of coefficients
- x_{1it} are observed predictor variables.
  - May not want to allocate activities strictly to involvement
- Identification:
  - $\alpha[1] = 1$ and if x includes intercepts, $\beta_{0i} = 0$

Model Part 2: Purchase Equation

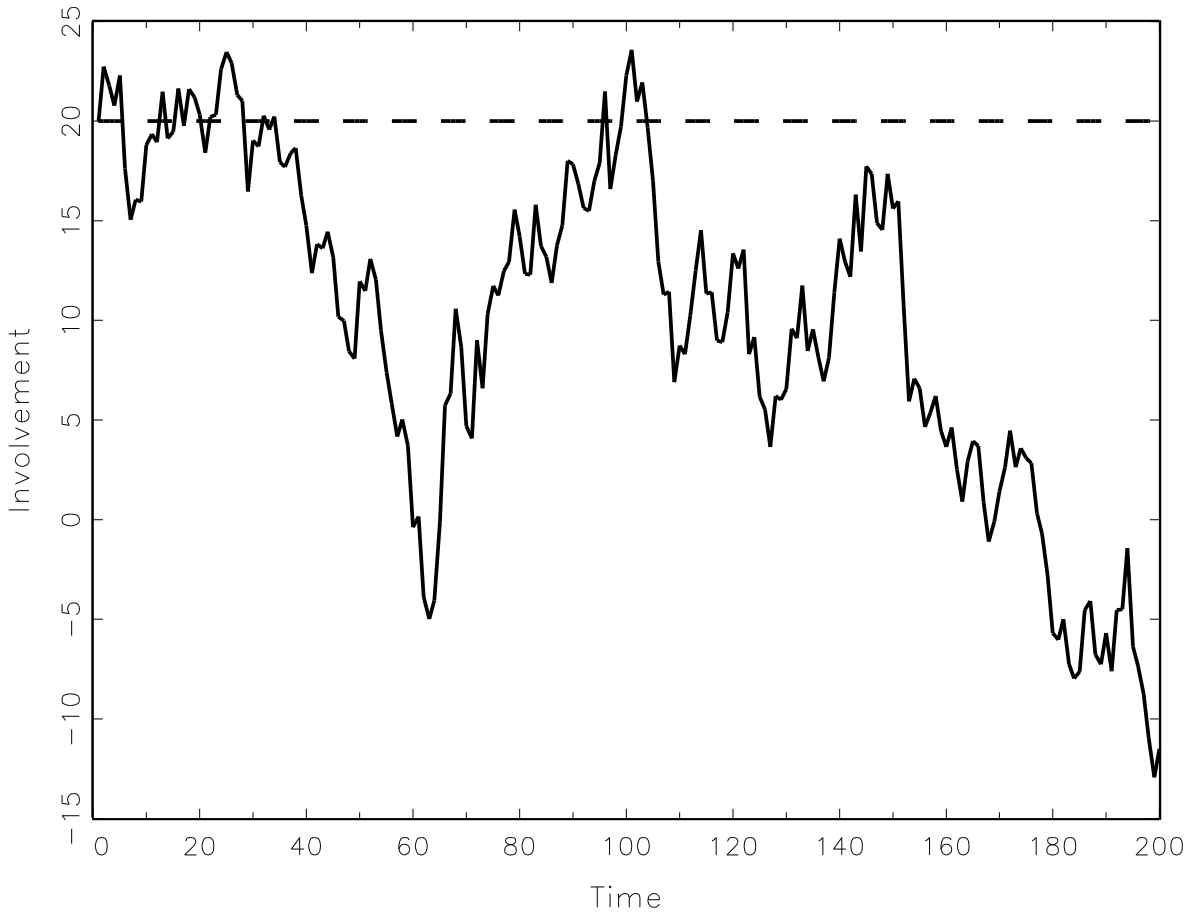
- Latent involvement predicts purchases
- $W_{it} = 1$ if user i makes a purchase at time t
- $W_{it} = 0$ if user i does not make a purchase at time t
- $P(W_{it} = 1) = \text{Probit}(\theta_{it})$

Model Part 3.a: Dynamic Model for Involvement

- $\theta_{it} = f(\theta_{i,t-1}) + \delta_{it}$ where $\delta_{it} \sim N(0, \nu^2)$
- Dynamic Linear Model: $\theta_{it} = G\theta_{i,t-1} + \delta_{it}$
- Random walk is most common choice in marketing
  - $\theta_{it} = \theta_{i,t-1} + \delta_{it}$
  - Works well for filtering
  - Nice interpretation based on sequential Bayes:
    - Prior at time t is the posterior at time t-1 plus a random disturbance
    - Random disturbance keeps posterior from concentrating on one value with long data series
  - Nonstationary model has infinite process variance
  - Allows involvement to wander everywhere

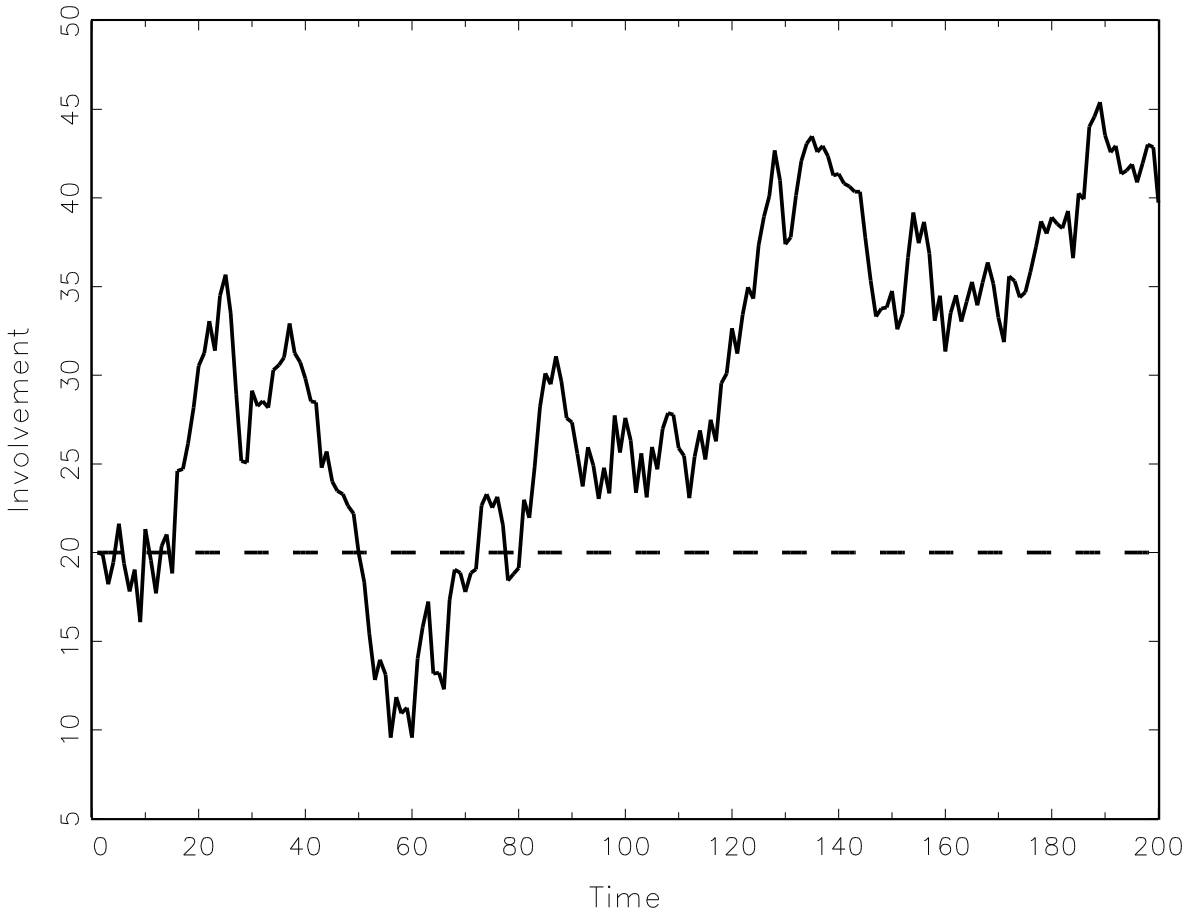
Random Walk: Realization 1

Start at 20; Error STD Dev = 2



Random Walk: Realization 2

Start at 20; Error STD Dev = 2



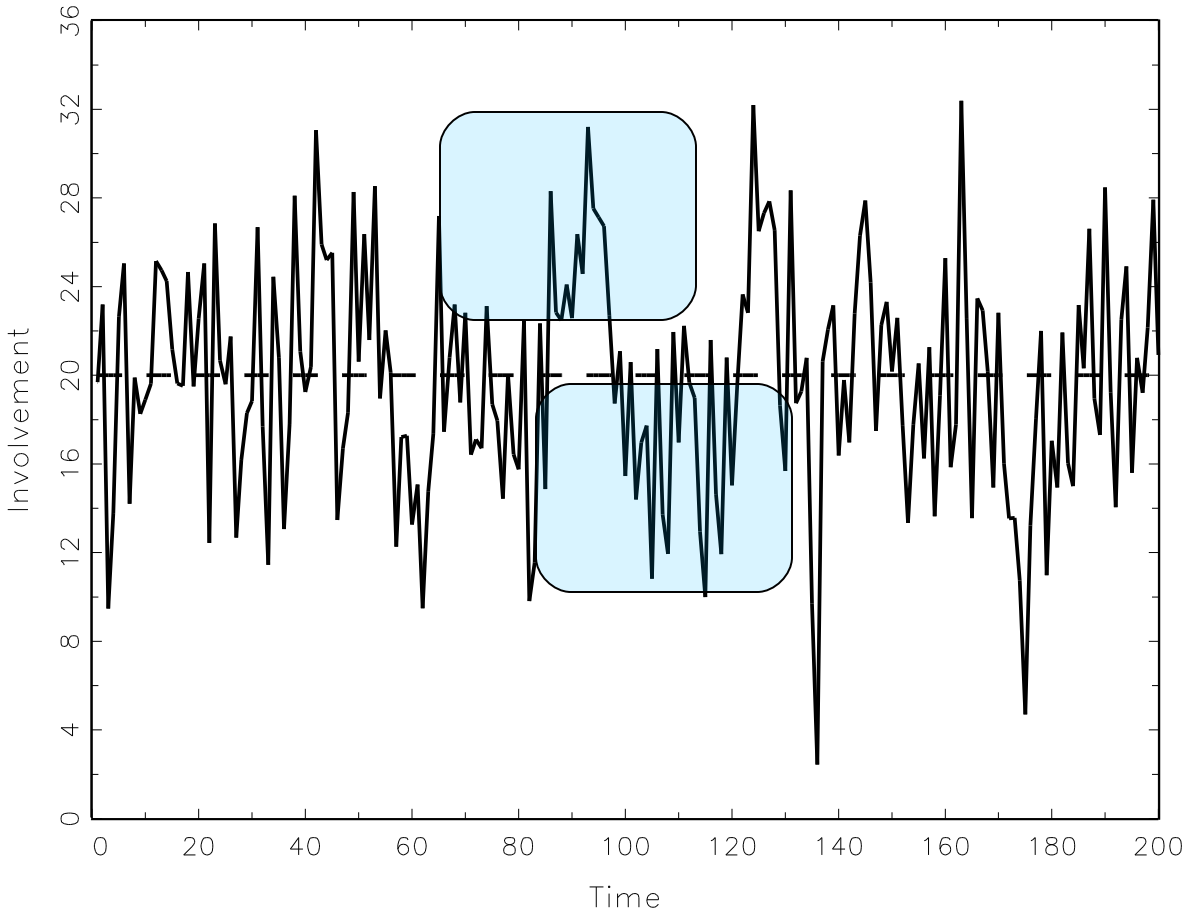
Mean Reverting, Stationary AR(1) Process

$$\theta_{it} = \phi + \psi(\theta_{i,t-1} - \phi) + \delta_{it} \text{ where } \delta_{it} \sim N(0, n^2)$$

- ψ is the AR(1) parameter
  - $|\psi| < 1$ is stationary condition
- ϕ is the process mean
  - Unconditional mean $E(\theta_{it})$
- $\nu^2/(1 - \psi^2)$ is process variance
  - Unconditional variance $V(\theta_{it})$
  - Process variance is finite, unlike random walk

AR(1) Process

Process Mean = 20; AR Parameter = 0.7; Error STD = 2



Series reverts to process mean

Periods above and below process mean

Process for Involvement

- AR(1) process is rather uninteresting for latent involvement
  - In the long-run, the involvement returns to the process mean
  - Random “shocks” cause involvement to deviate from process mean
  - Some “stickiness” due to AR model
- We want to know if customers progress up an involvement ladder, not stay at a constant level

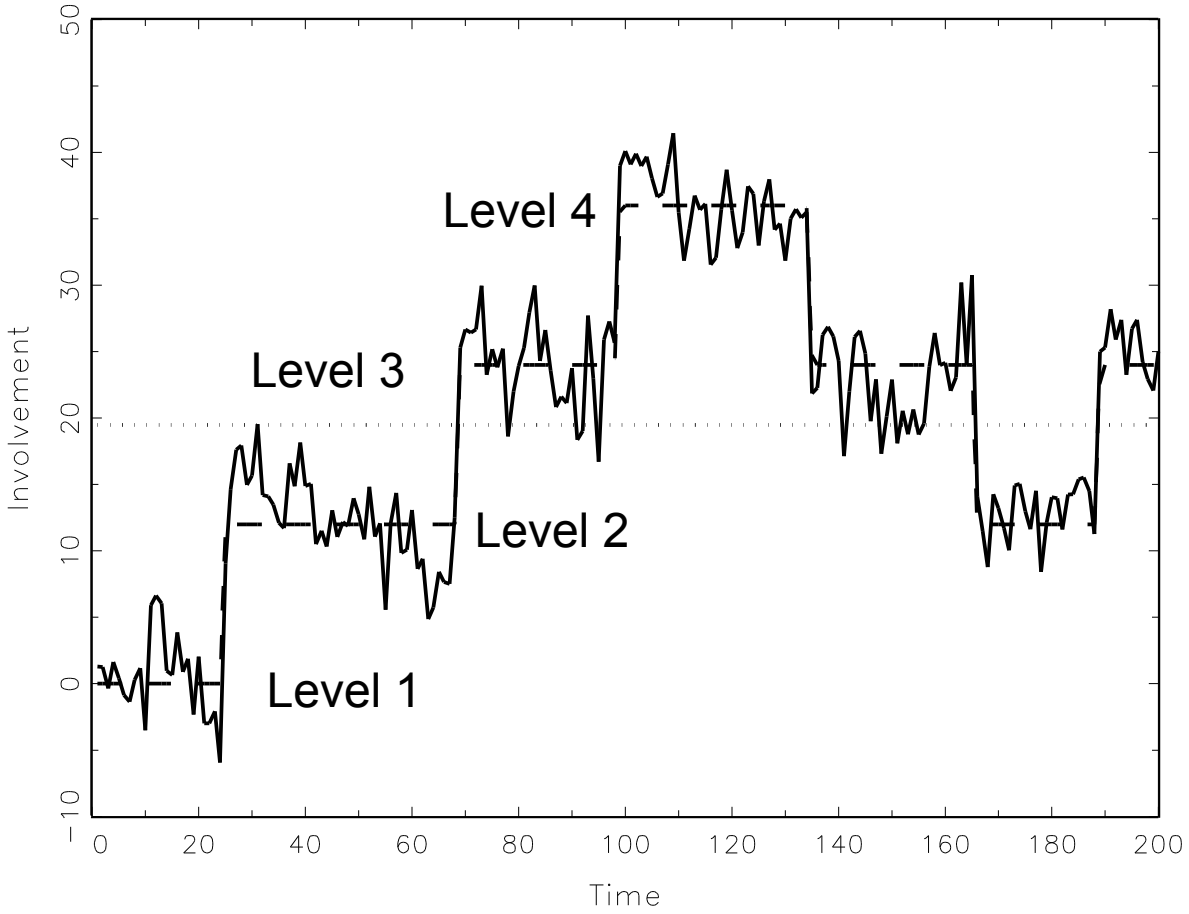
Model Part 3.b: Dynamic Segmentation of Involvement

$$\theta_{it} = x'_{2it}\varphi_s + \psi\left(\theta_{i,t-1} - x'_{2it}\varphi_s\right) + \delta_{it} \text{ and } \delta_{it} \sim \mathrm{N}\left(0, \nu^2\right)$$

- Mean reverting AR(1) process
- s = s(i,t):
  - Subject i belongs to segment s at time t: s = 1, …, S.
- $x'_{2it}\varphi_s$ is long-term process mean for segment s
  - Intercept ϕ_{0s}
- ψ is AR parameter
- $\lambda_{i,s}$ = P(Segment s), user specific

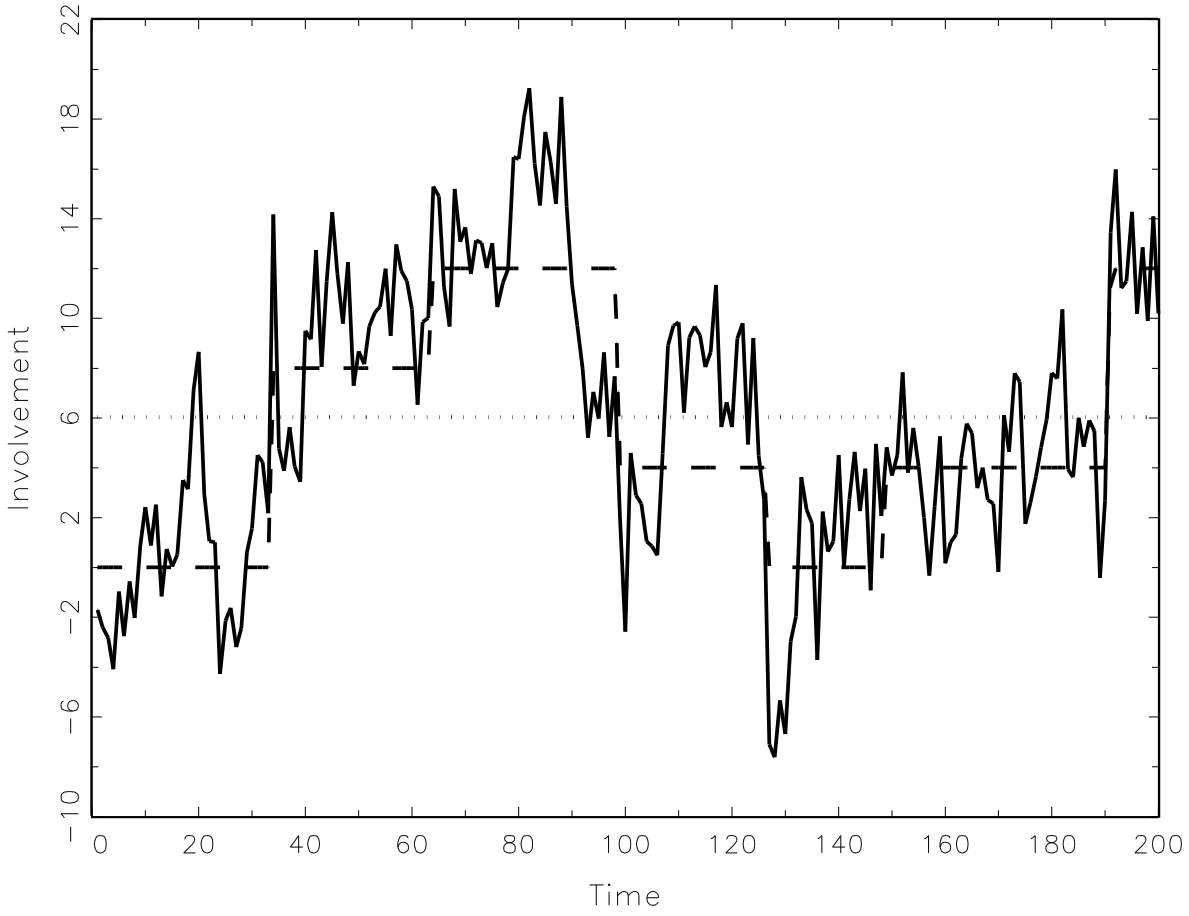
Good Discrimination of Segments

4 or more error standard deviations between levels



Poor Discrimination

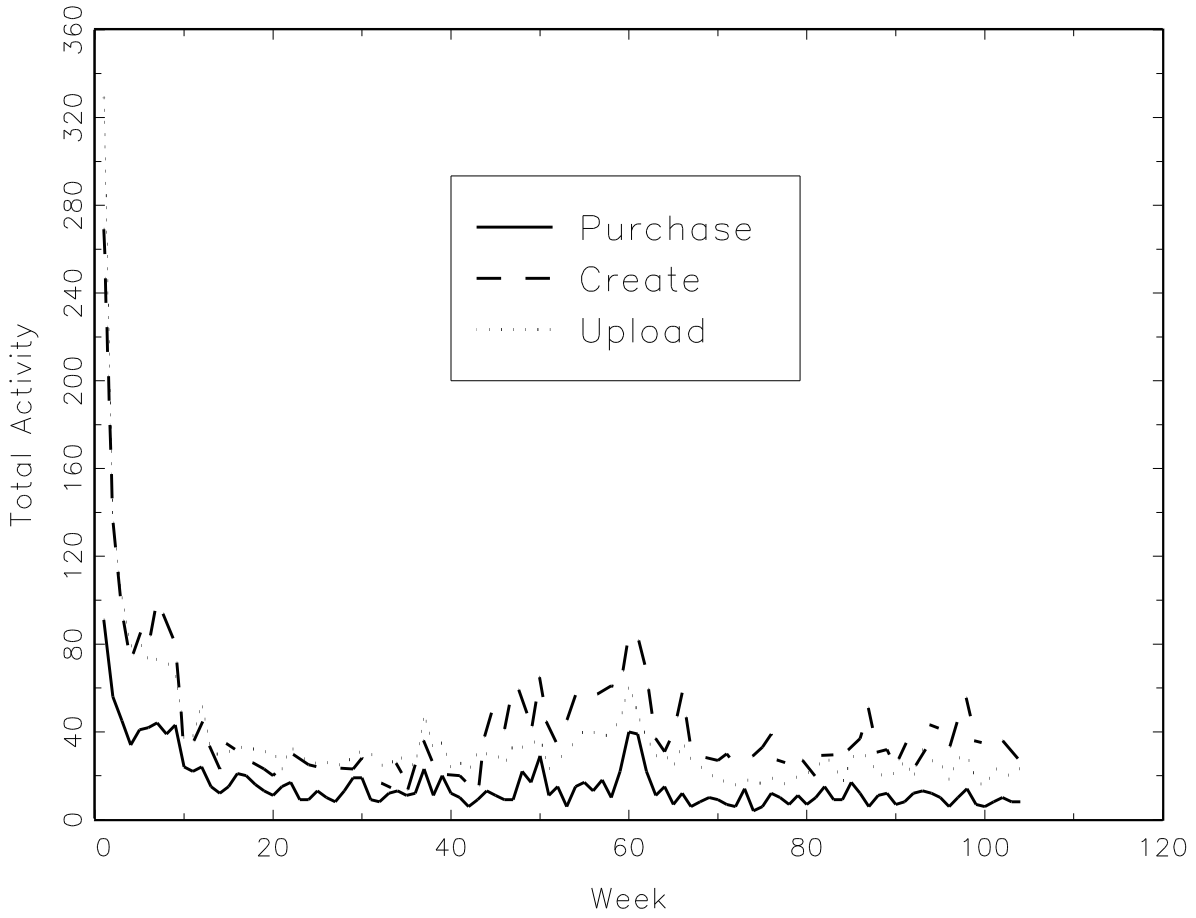
Estimates one segment with inflated AR and Error STD DEV



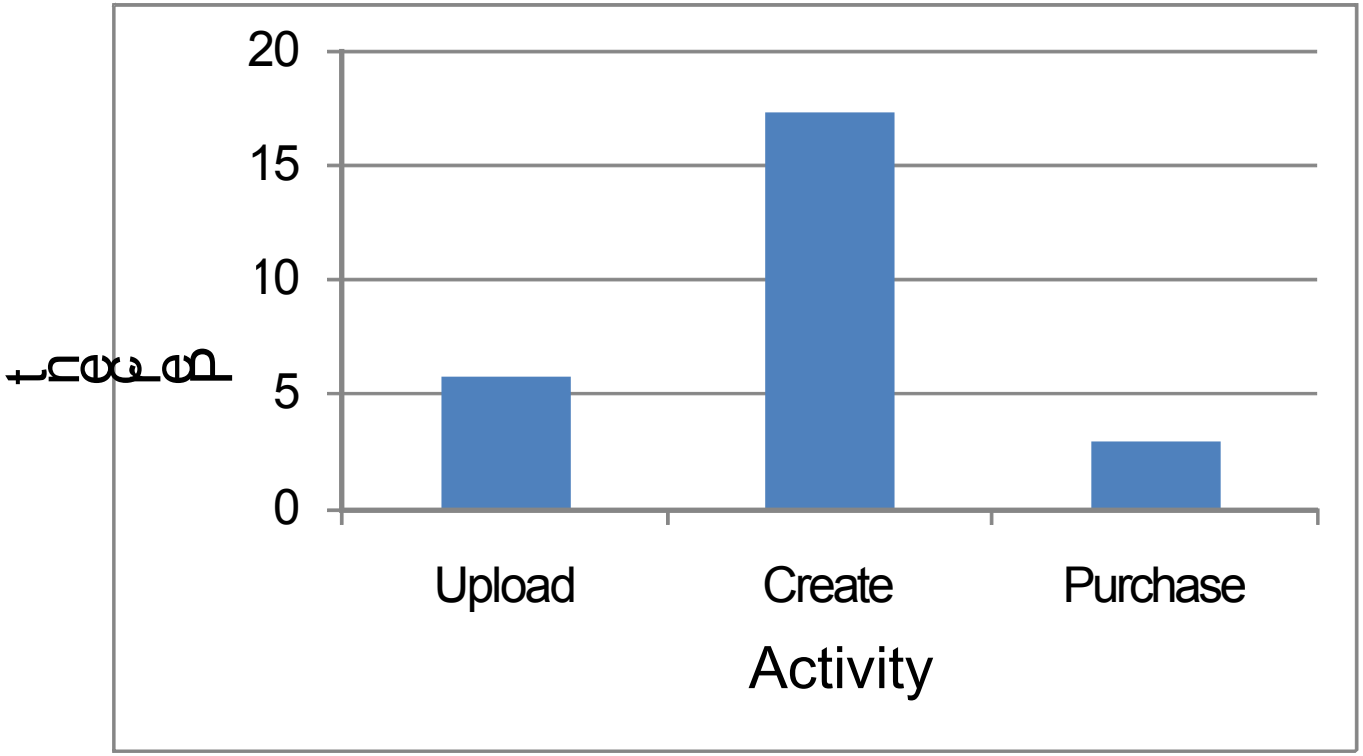
Identifying Segments

- When segments are well separated, order probabilities of segment membership
- When segments are not well separated, force segment intercepts to fall within specified bands
  - Pick bands to satisfy managerial objectives
  - Number of segments
  - Width of bands
  - Need to know likely range of $\{q_{it}\}$
  - Some trial and error (!)

Total Weekly Activity

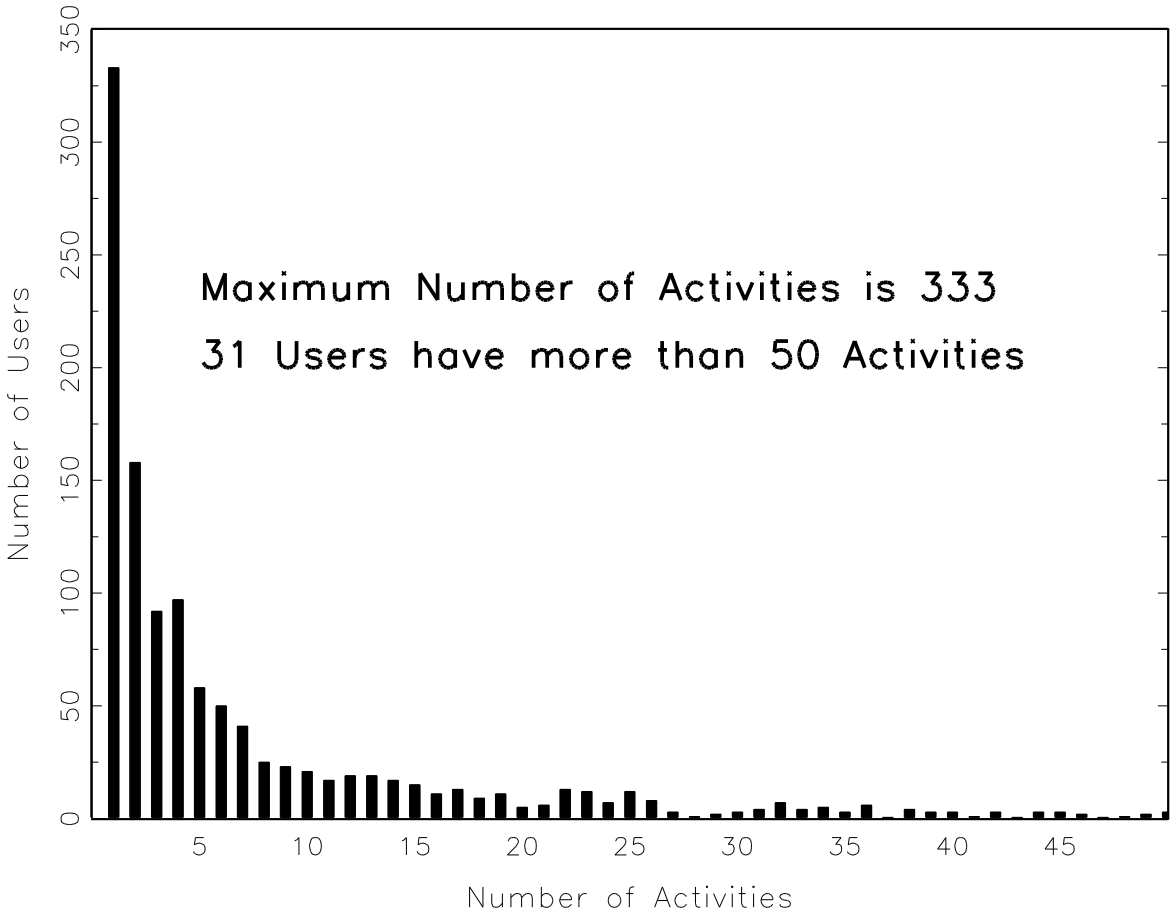


Activity by Weeks and Users



Long Tail for Activities

All users over 104 weeks

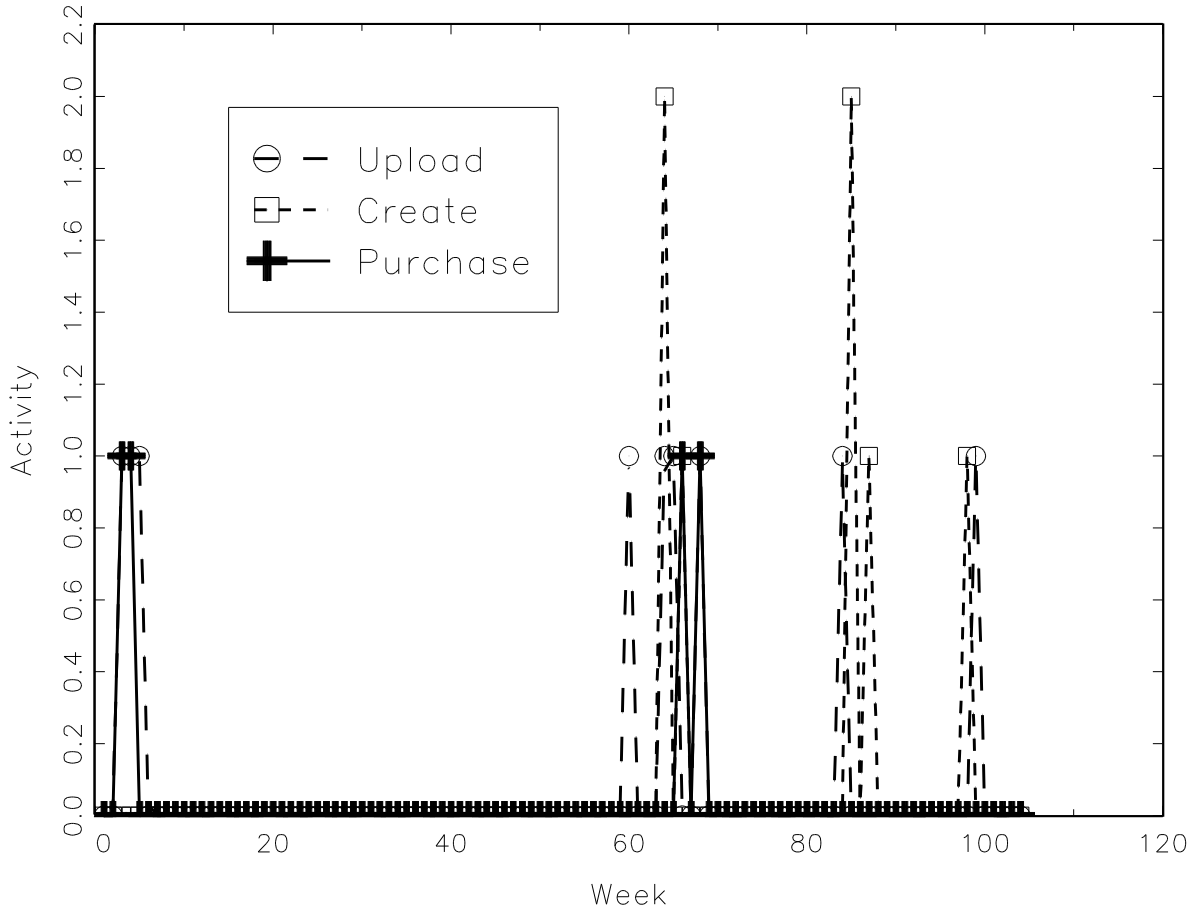


Bursty Stuff

- Enthusiasts in the long tail consistently use site on a regular basis
- Other users have periodic and aperiodic cycles
- Probably triggered by motivating events
  - Holidays
  - Reunions or get-togethers
  - Birthdays
  - Etc.

A Fairly Active Person

Above 95th Percentile in Total Activity



Purchases, if they occur, often follow Uploads and Creates

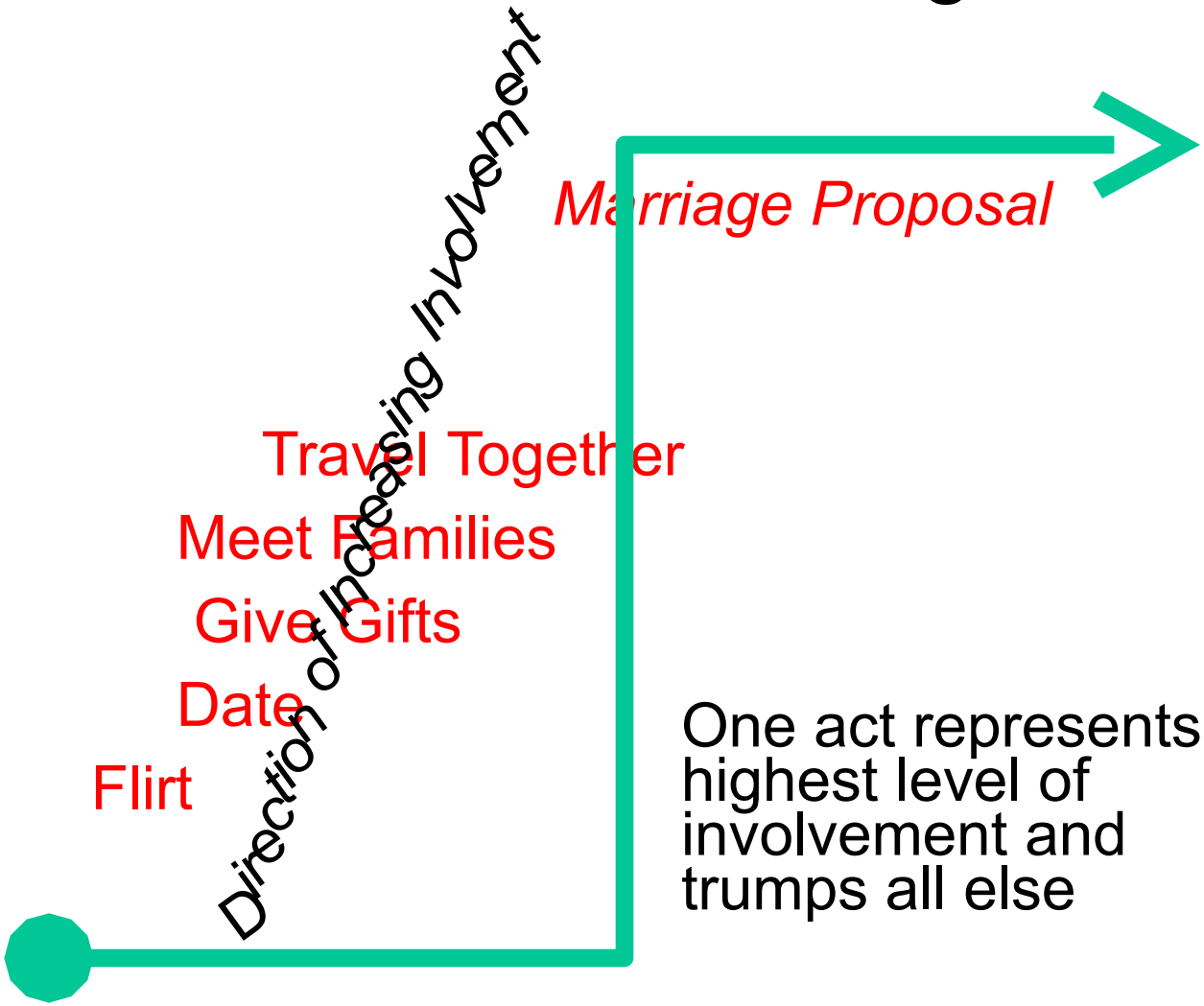
Kernel Smoother

- Time of kth event of type j for subject i: ti,j,k
- Number of events of type j at time ti,j,k: ni,j,k
- Kernel Smoother:

$$W_{ijt} = \sum_{s=1}^{N} n_{ijk} K\left(t, t_{ijk}\right)$$

- Not normalized to be a density estimator
- Use normal kernel with standard deviation = 4 weeks
- Outstanding issue …. Involvement laddering

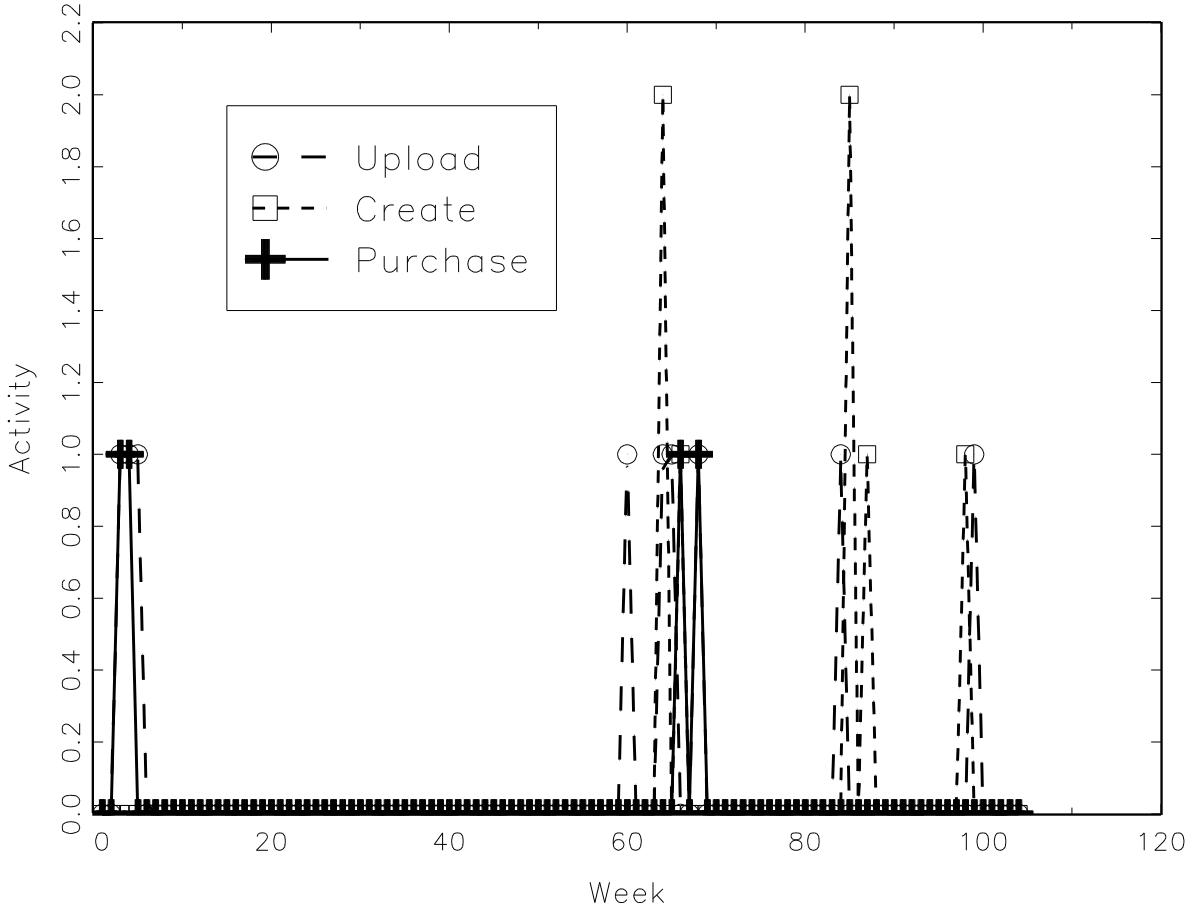
Involvement Laddering



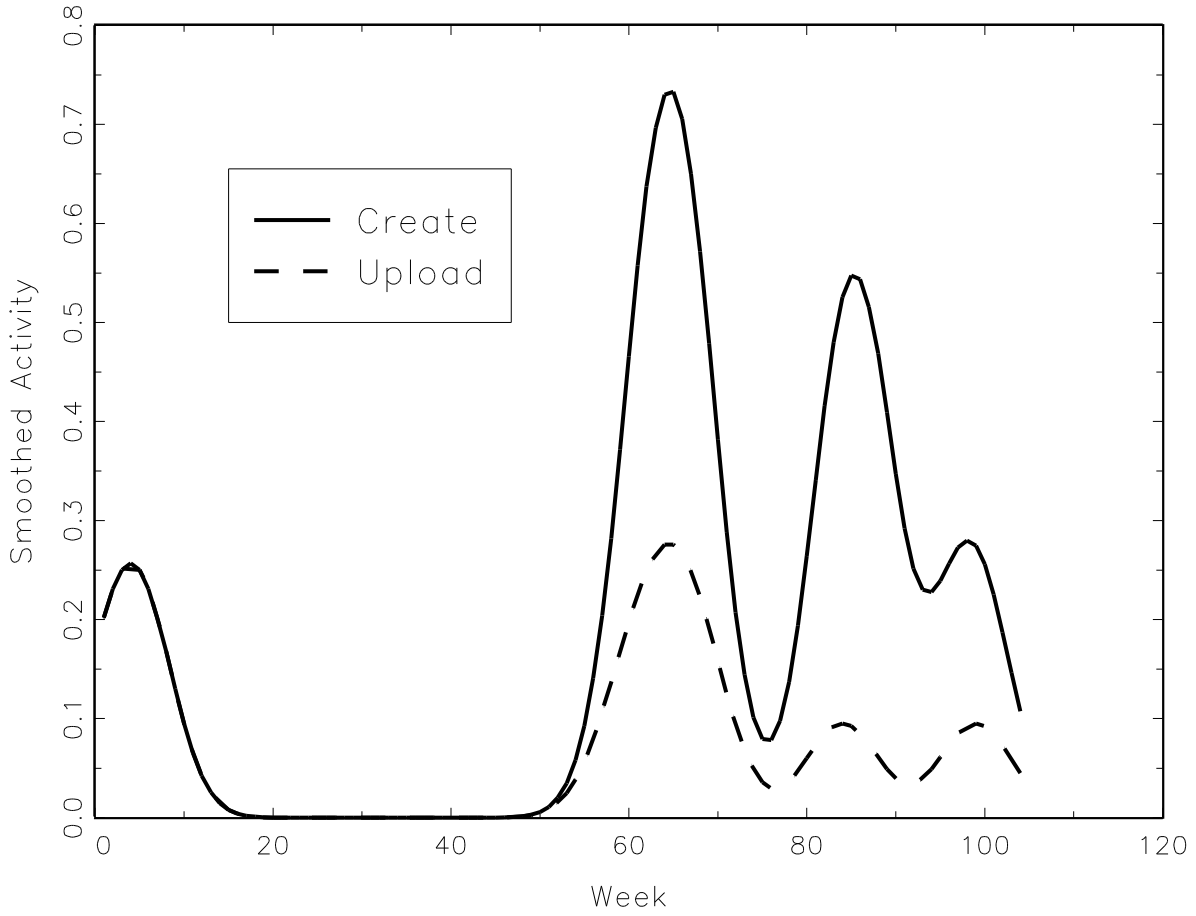
Our Data

- Creating content represents higher level of involvement than uploading photos
- Need to scale data so that creating content is larger than uploading in order to have positive relation to latent involvement
- Cumulative product of kernel smoothers in order of involvement

Observed Activity for User X



Multiplicative Data Smoother



Results: Observational Equations

Alpha	Post Mean	Post STD DEV
UPLOAD	1.000	0.000
CREATE	2.848	0.008
PURCHASE	**1.505**	**0.054**

Intercepts	Post Mean	Post STD DEV
UPLOAD	0.000	0.000
CREATE	0.008	0.001
PURCHASE	**-0.880**	**0.025**

Error Covariance: Posterior Mean

Sigma	UPLOAD	CREATE	PURCHASE
UPLOAD	0.00206	-0.00072	0.00231
CREATE	-0.00072	0.00109	-0.00082
PURCHASE	0.00231	-0.00082	**0.15463**

Error Covariance: Posterior STD DEV

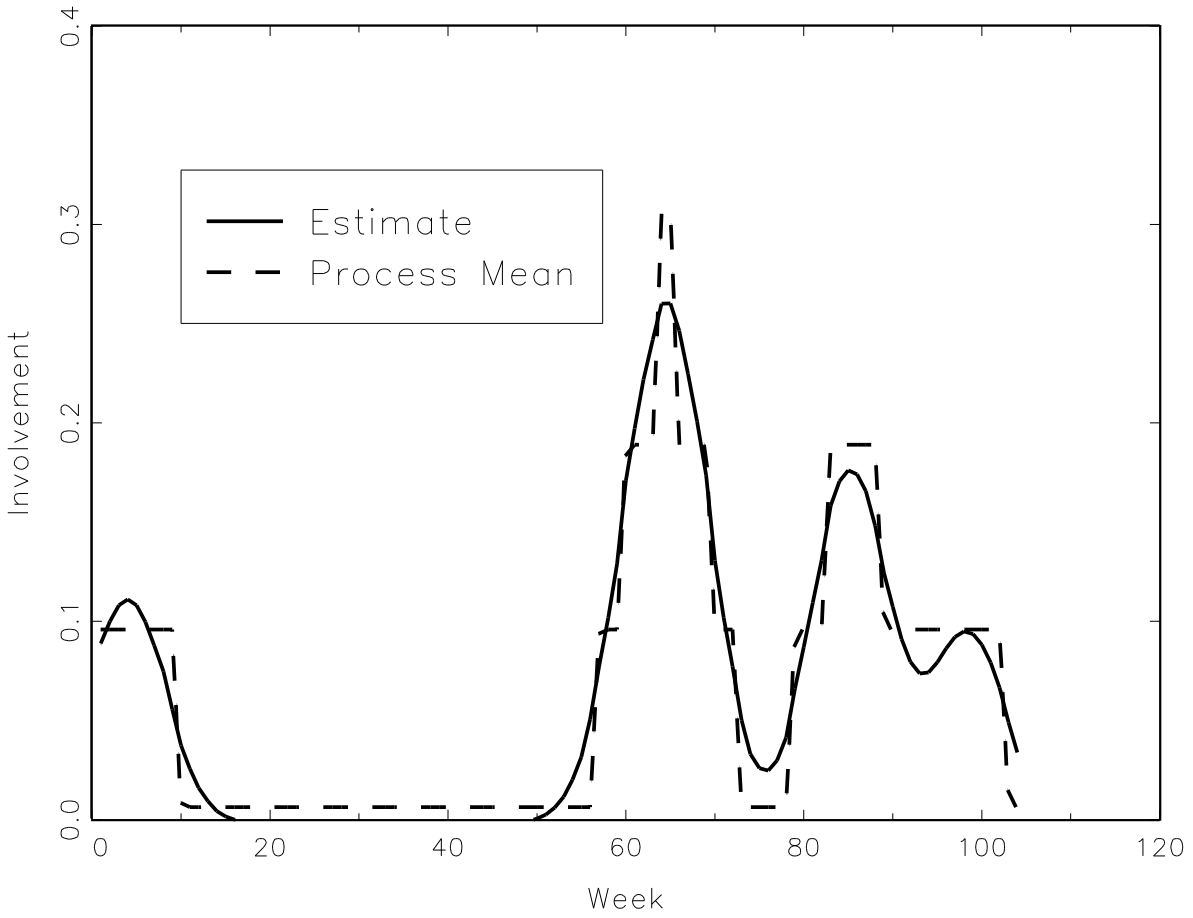
Sigma	UPLOAD	CREATE	PURCHASE
UPLOAD	0.00008	0.00009	0.00020
CREATE	0.00009	0.00012	0.00022
PURCHASE	0.00020	0.00022	**0.00872**

Probit model for Purchases

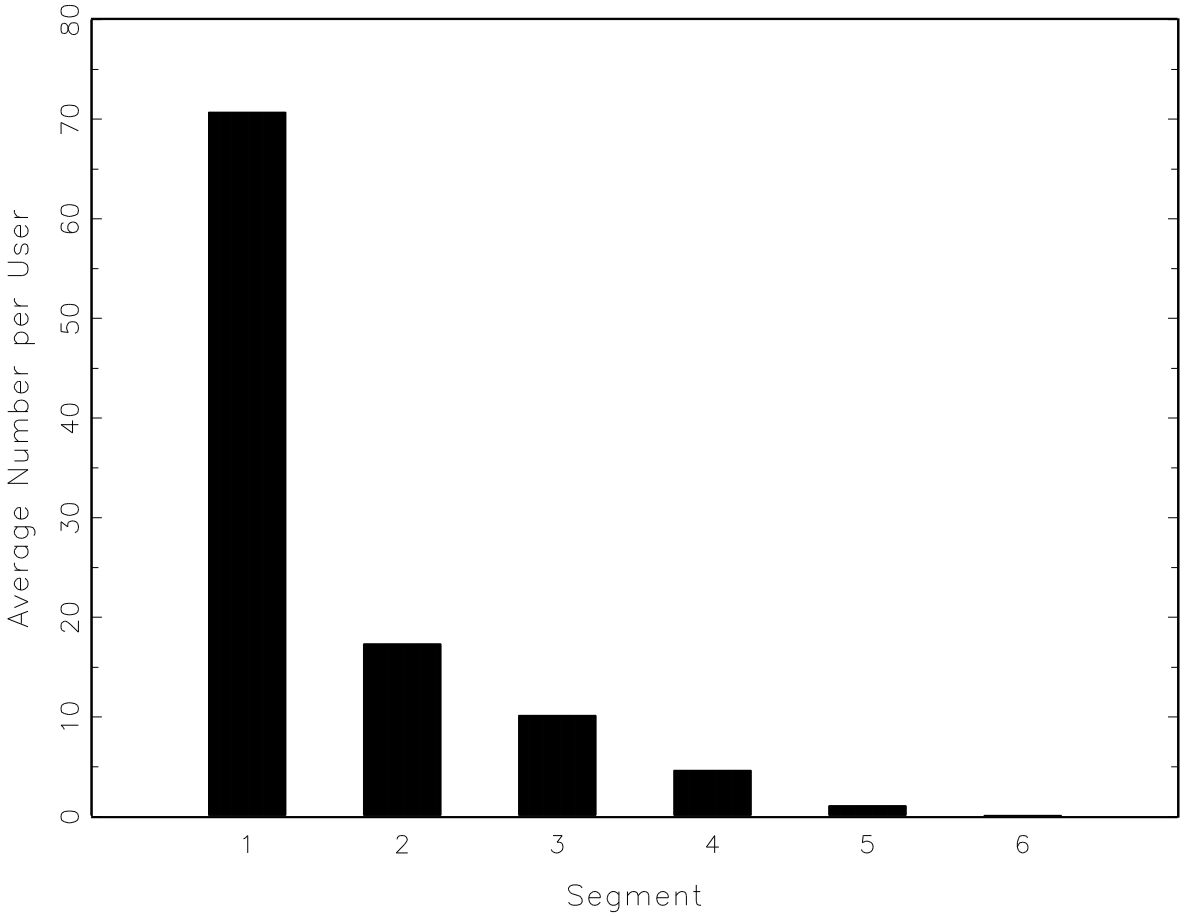
Results: Dynamic Involvement Model

Segment	Post Mean	Post STD DEV
1	0.006	0.0003
2	0.096	0.0006
3	0.189	0.0008
4	0.306	0.0011
5	0.501	0.0009
6	0.846	0.0264
PSI AR(1)	0.379	0.0097
NU Error STD	0.023	0.0001

Estimated Involvement User X



Segment Classifications
Averaged over Subjects: 104 Weeks



Follow-ups: Implementation and So On

- Marketing mix would be most useful!
  - Marketing programs, user interface changes, other potentially important covariates would be interesting
- Compare to RFM metrics
- Feed results into CRM
- Follow up primary research with high- and low-involvement customers
- Explore other data models
  - Compounded Poisson processes may be useful
  - non-/pseudo parametric techniques
  - Nonlinear dynamical systems application?
  - Simpler parametric specifications