



# **We know what you did at 9am**

## **Analysis Systems with Dynamic User Generated Content**

Christian Wallenta, Oxford  
Mohamed Ahmed, UCL  
Ian Brown, Oxford  
Stephen Hailes, UCL  
Felipe Huici, UCL

Multi Service Networks 2008  
10.07.2008

# Motivation

- Understand how *data enters* these systems
- Understand how *data evolves* over time
  - > Derive models that explain *when* and from *where* data comes into these systems
  - > *Apply these models* to a wider range of applications to optimise their performance


# Quick Overview


**digg™** Join Digg | About | Login


Technology ▾ World & Business ▾ Science ▾ Gaming ▾ Lifestyle ▾ Entertainment ▾


Popular Upcoming News Videos Images Podcasts Custom

News, Images, Videos — **Most Recent** Top in 24 Hr 7 Days 30 Days 365 Days



**139 diggs**  
[Obama Supporters Take His Middle Name as Their Own](#)  
nytimes.com — A growing band of supporters of Obama are expressing solidarity with him by informally adopting his middle name, Hussein. [More...](#) (US Elections 2008)  
[digg it](#) 59 Comments [Share](#) [Bury](#)  moakb made popular **6 min ago**

**247 diggs**  
[10 Sci-Fi Books That Were Better Off on Paper](#)  
io9.com — Here are 10 books that we think should never have been committed to celluloid. [More...](#) (Educational)  
[digg it](#) 44 Comments [Share](#) [Bury](#)  msaleem made popular **14 min ago**

**193 diggs**  
[In-Depth look at the DIY "Bucket Hydroelectric Generator"](#)  
aidg.org — Pico Hydroelectric Generator in 5 gallon bucket developed by Sam Redfield and tested at La Florida in Guatemala. The generator is meant to be a very small, cheap, low impact generator designed to be used with existing gravity fed irrigation, fresh water, or waste water systems. [More...](#) (Environment)  
[digg it](#) 14 Comments [Share](#) [Bury](#)  mark076h made popular **31 min ago**

**reddit** what's hot new controversial top

↑ Vote up if you think going to war with Iran is a bad idea (self.reddit.com)  
1 2461 submitted 13 hours ago by addie25 to reddit.com  
↓ 363 comments

↑ Awkward [PIC] (photobaseement.com)  
2 253 submitted 2 hours ago by jda06 to pics  
↓ 58 comments

↑ No, THIS is a real war hero (en.wikipedia.org)  
3 1093 submitted 12 hours ago by anarchistica to politics  
↓ 153 comments

↑ House Resolution Calls for Naval Blockade against Iran. The measures called for in the resolutions amount to an act of war. (globalresearch.ca)  
4 122 submitted 3 hours ago by Escafane to worldnews  
↓ 43 comments

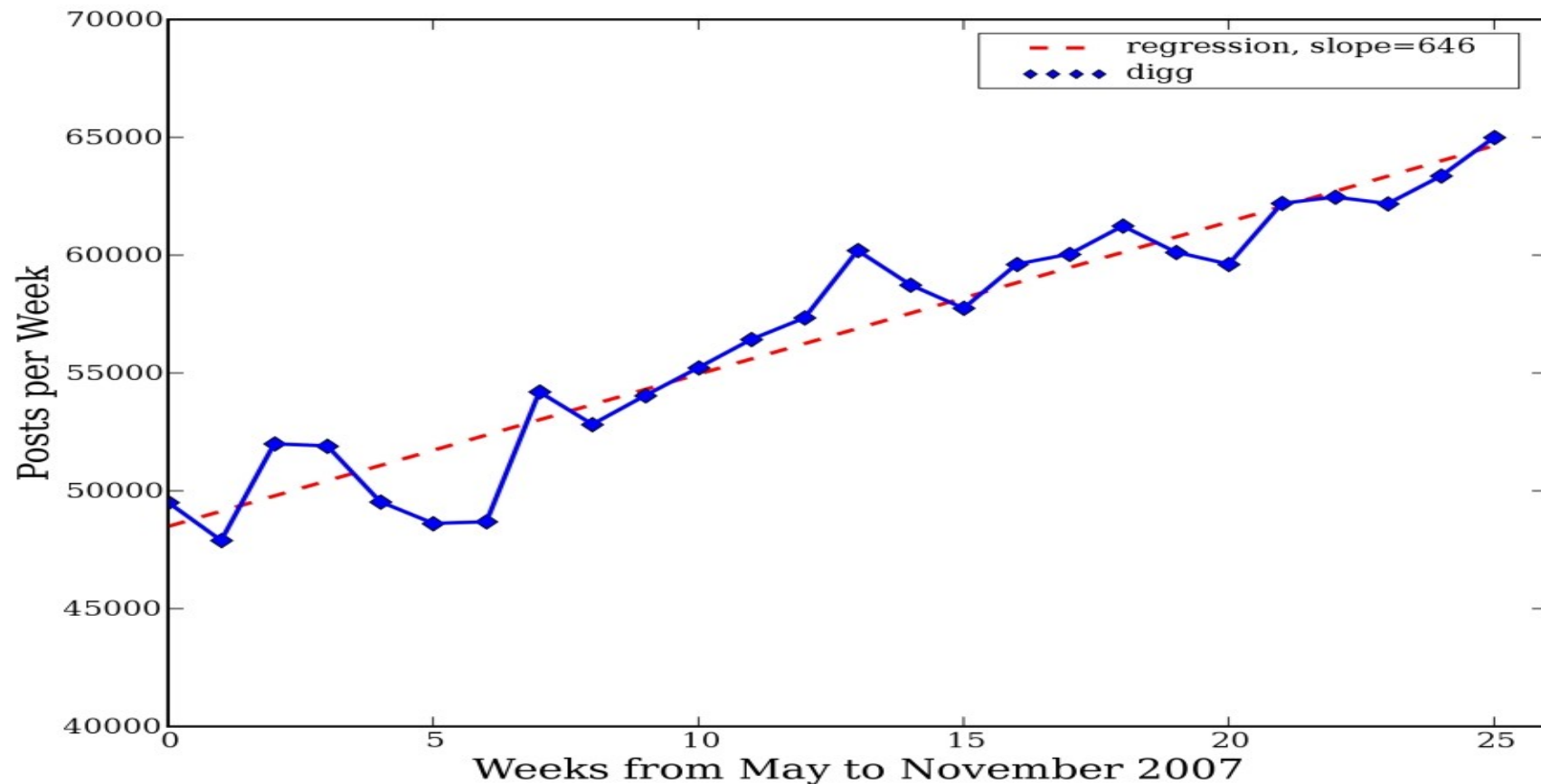
↑ The great ocean migration... Measuring up to 6' 6" across, poisonous golden cow-nose stingrays migrate in groups, called 'fevers', of up to 10,000 (Amazing pictures) (dailymail.co.uk)  
5 161 submitted 4 hours ago by allie to science  
↓ 24 comments

# Datasets

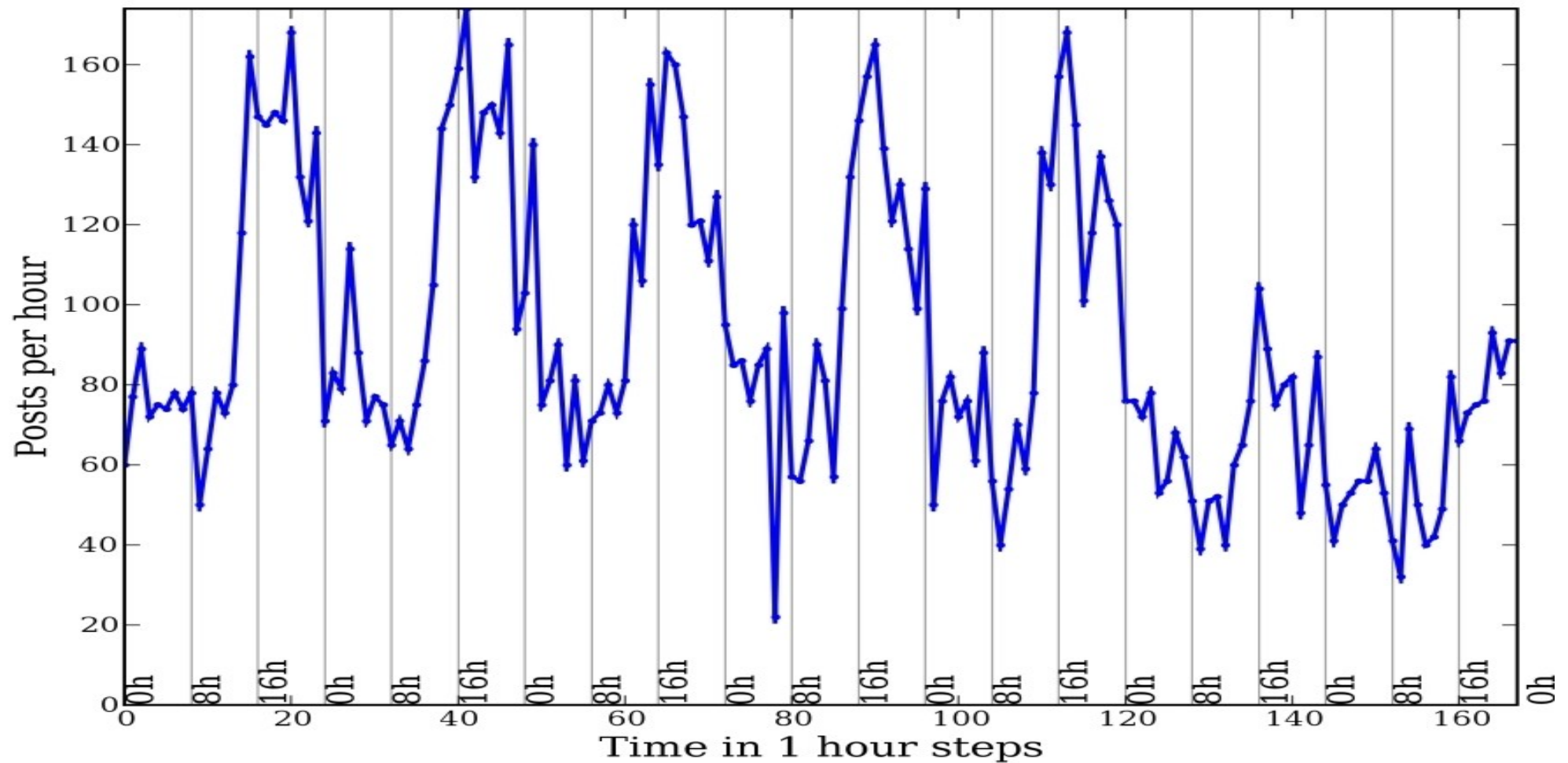
- *digg.com*
  - 1.5 million *posts* including submission time, author, number of votes between May and November 2007
  - 1.6 millions *votes* for 87,000 posts between Nov, 21<sup>st</sup> and Dec, 1<sup>st</sup> 2007
  - 240,000 *user profiles*
- *reddit.com*
  - 183,000 *posts* (Nov 07 to Feb 08)
  - 13,300 posts + *votes* (Nov, 23<sup>rd</sup> to Nov, 30<sup>th</sup>)

# Content Generation Trend

50,000 posts in May to 65,000 in November 2007

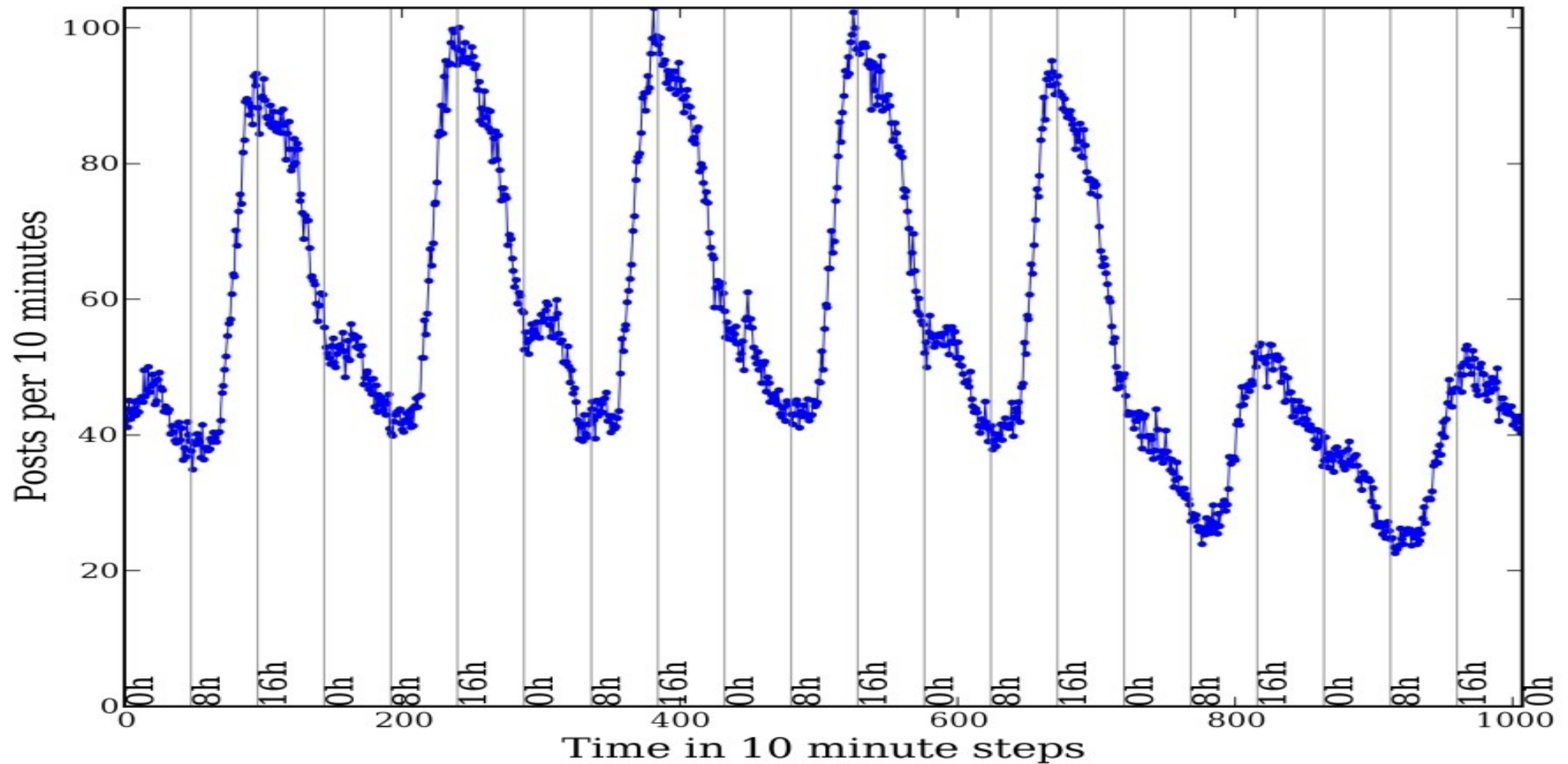


# Content Generation Volume per Week



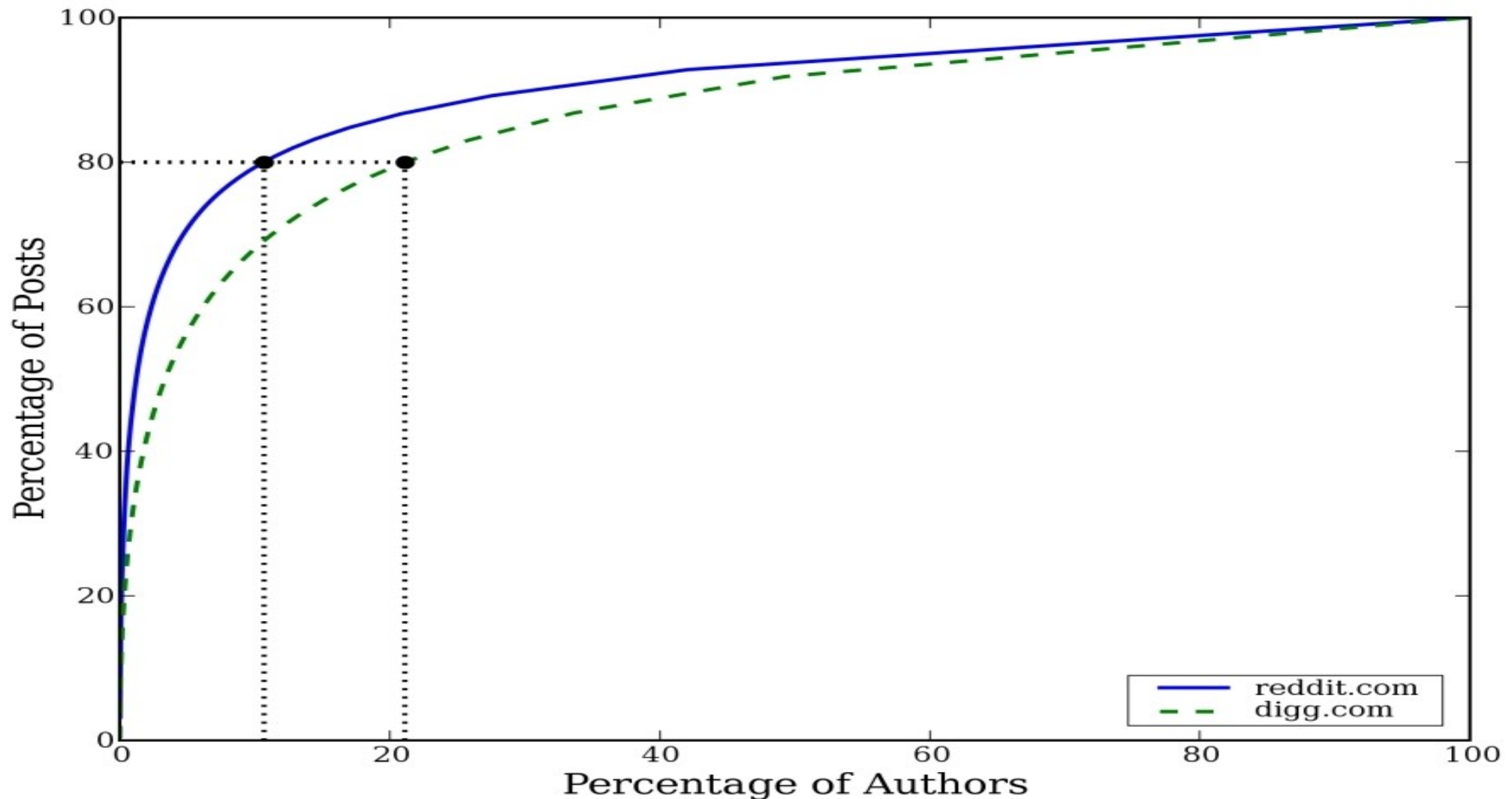
reddit.com

# Content Generation Volume per Week



digg.com

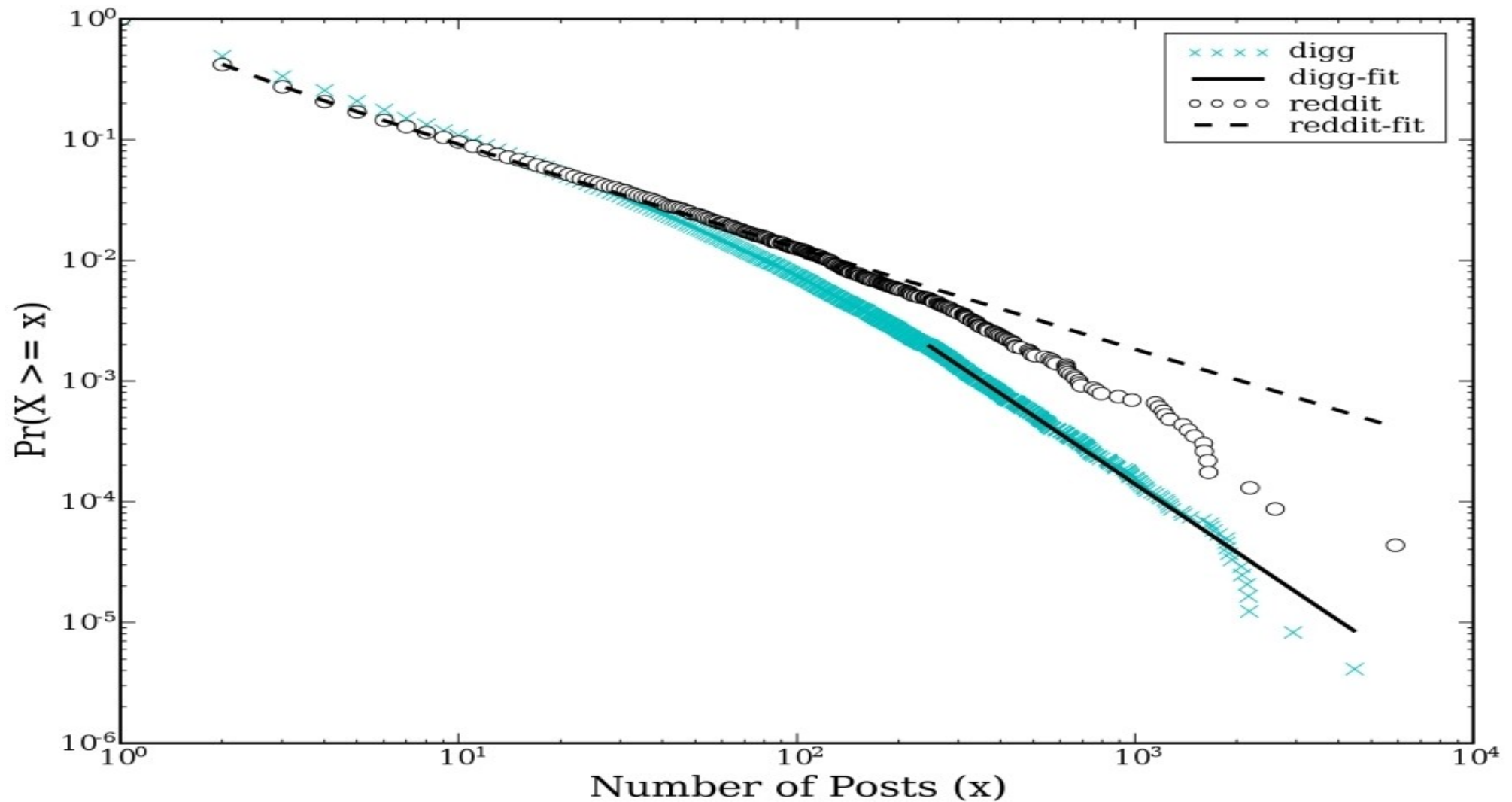
# Content Generation User Contribution





# Content Generation

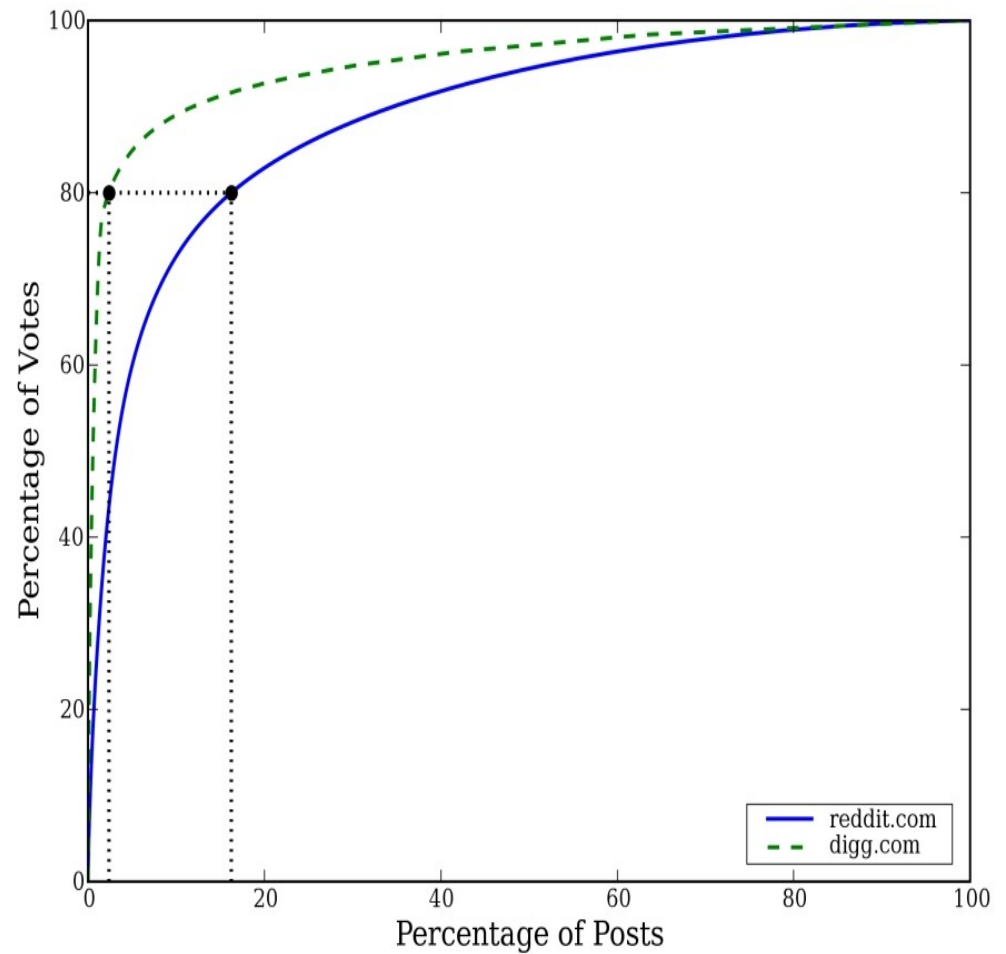
## User Contribution



# Popularity Analysis

## Votes Distribution

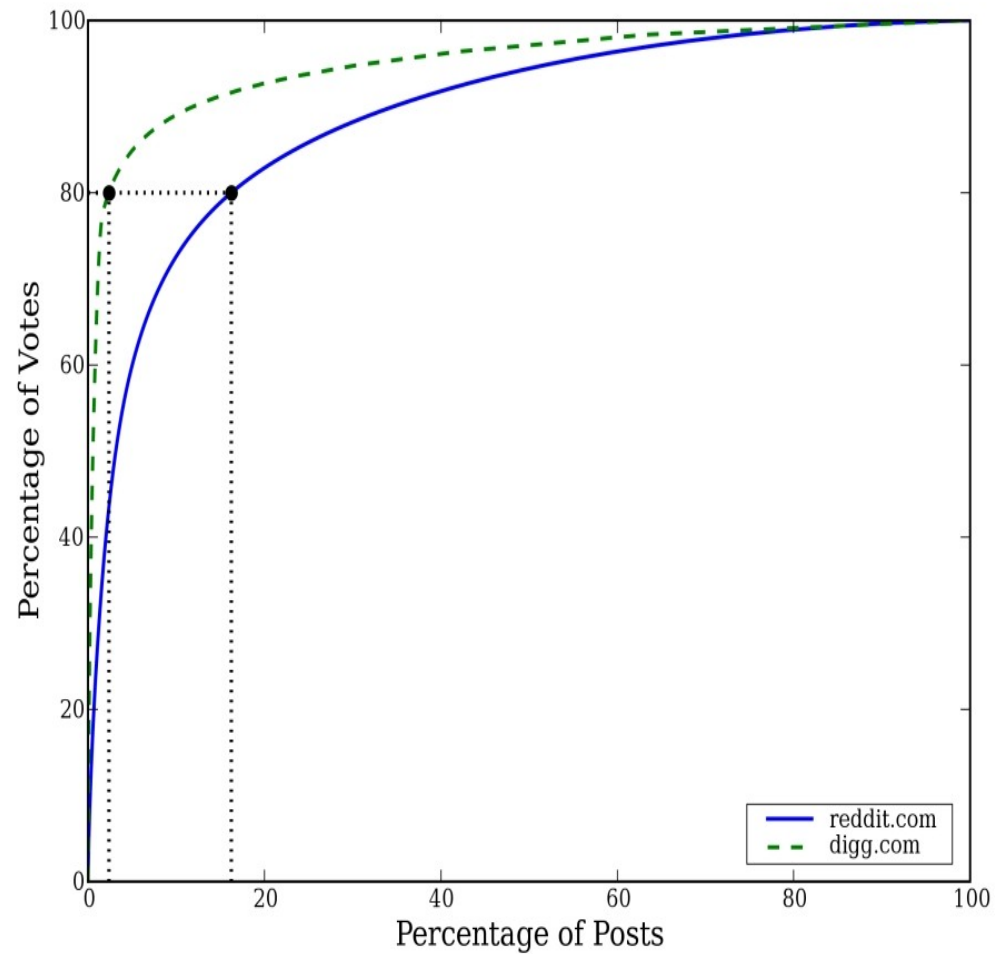
- What % of the votes goes to what % of the post?



# Popularity Analysis

## Votes Distribution

- What % of the votes goes to what % of the post?
- If votes ~ popularity then this distribution is always interesting for caching



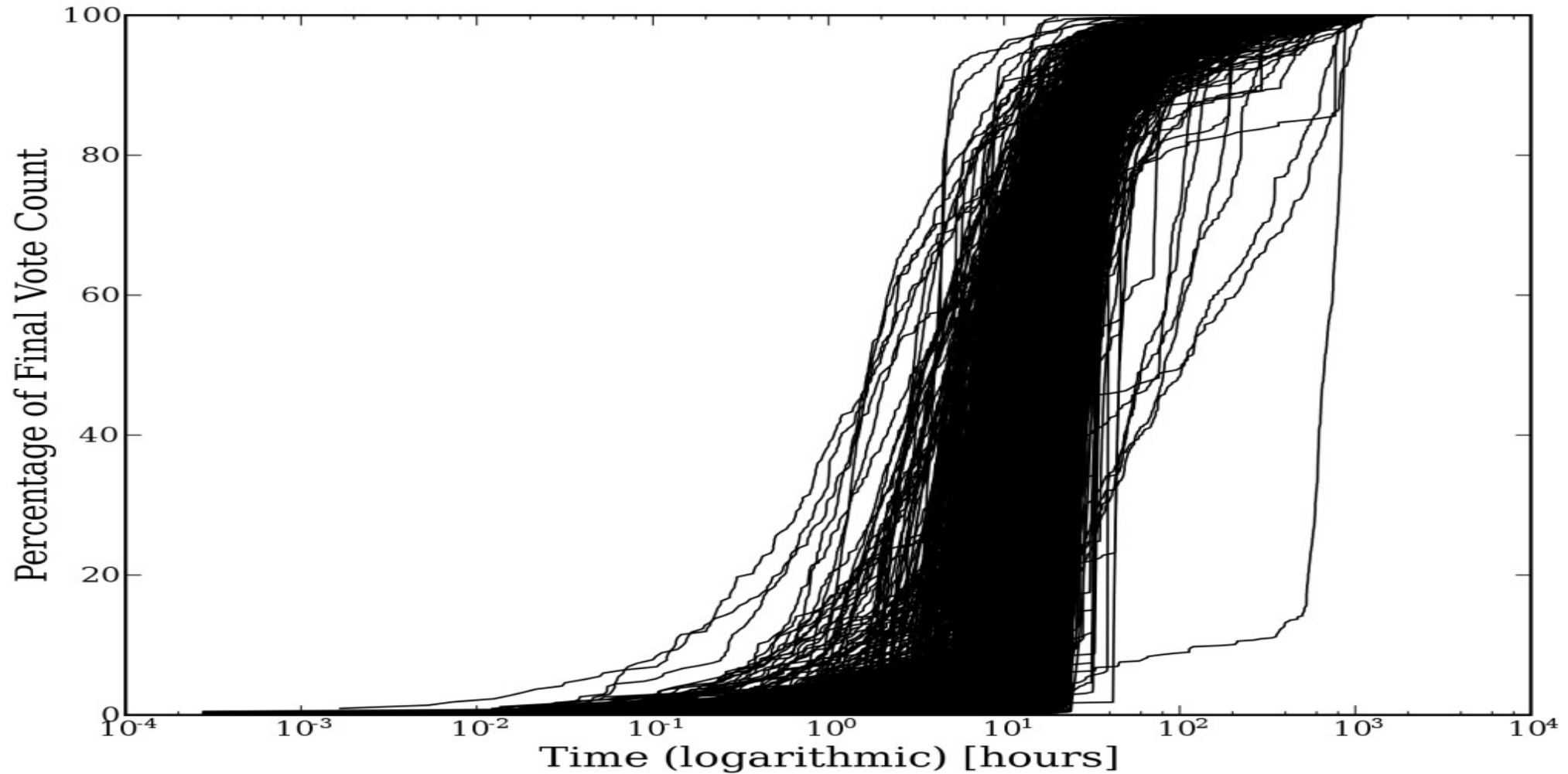
# Popularity Analysis

## Popularity Evolution

- Now we know static behaviour, but...
- How fast does this happen?
- How long does content stay popular?
- Monitor posts from submission time until they become inactive

# Popularity Analysis

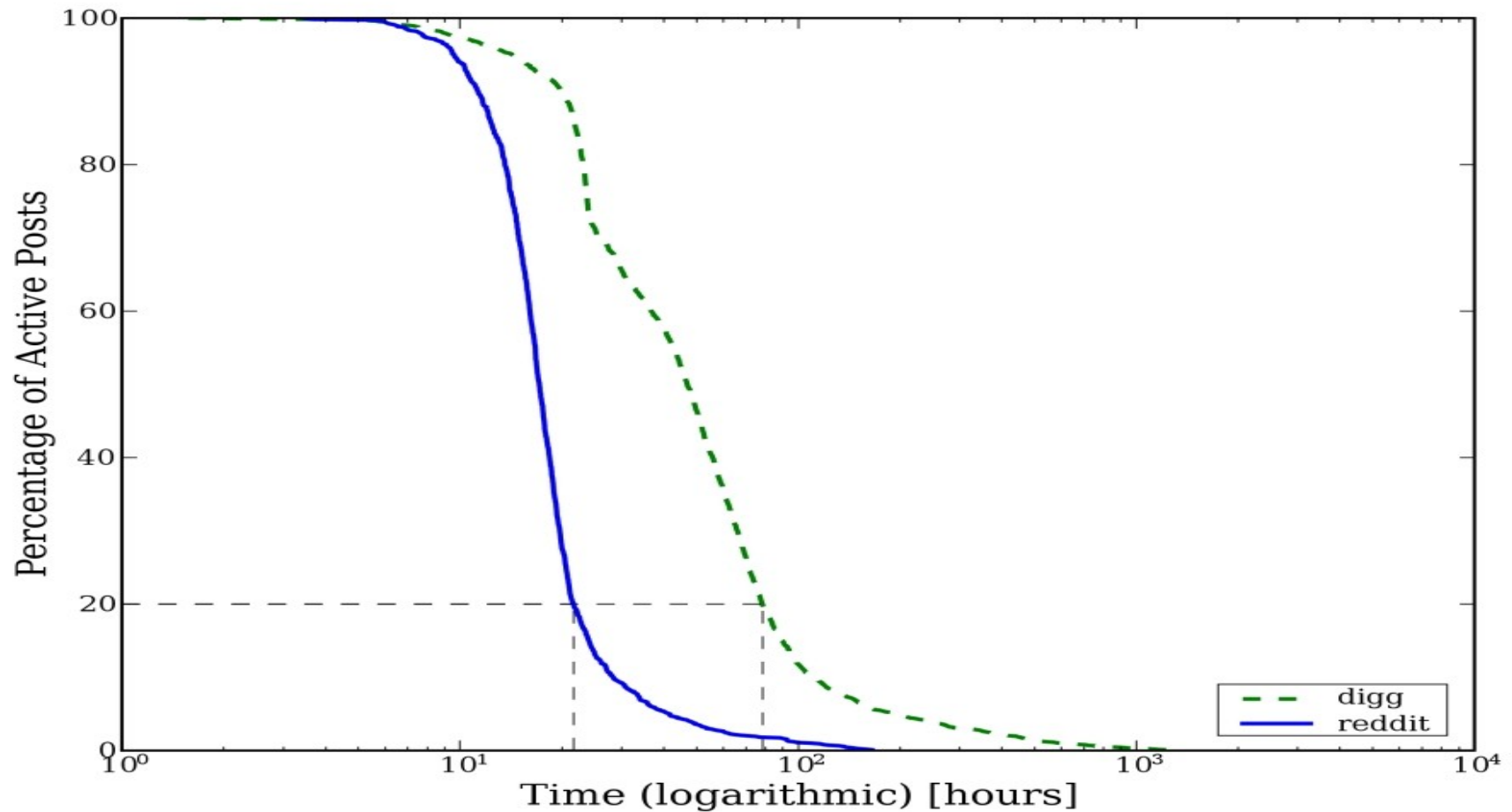
## Popularity Evolution



digg.com

# Popularity Analysis

## Post Lifetime



# Analysis Summary

- Lots of content, periodic patterns
- Few users create most of the content
- Most votes go to a few posts
- Content becomes popular fast, and has a short lifetime in contrast to e.g. YouTube

# Data Generation Model

## Motivation

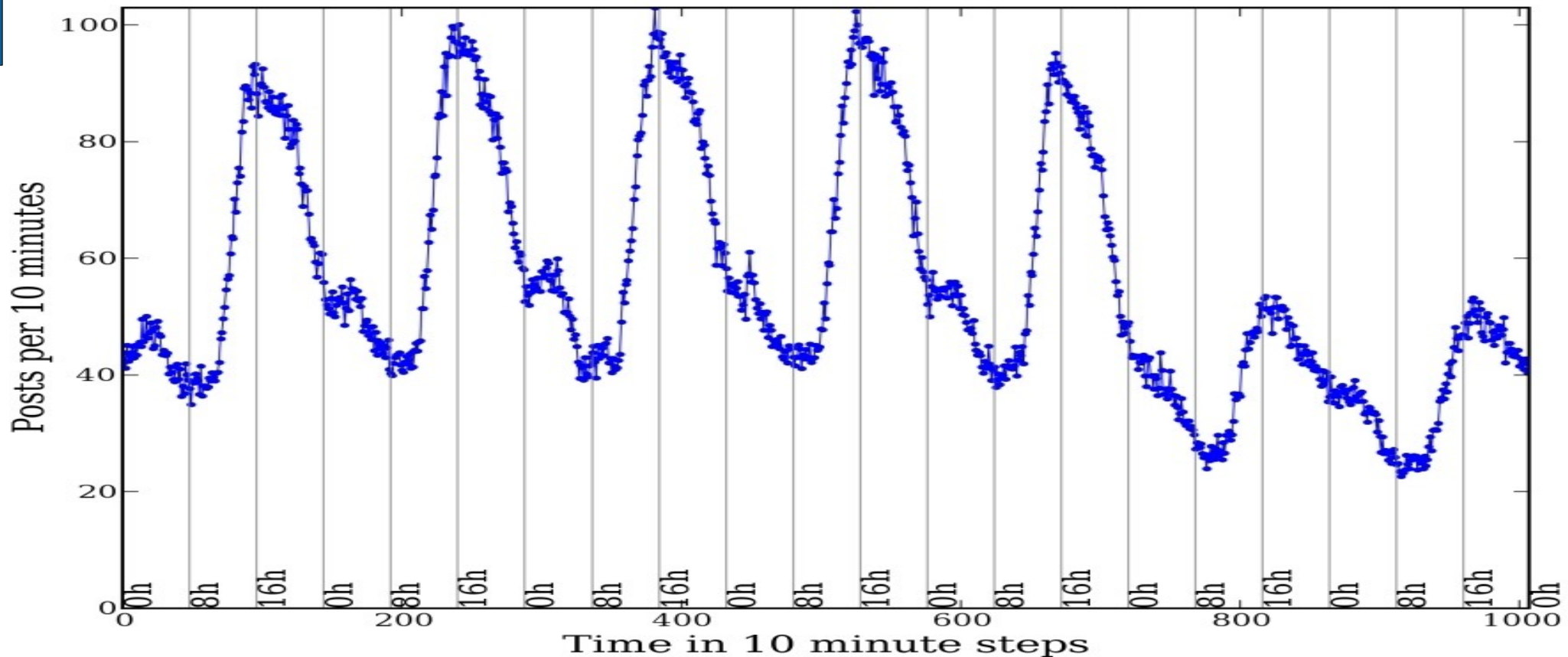
- Understanding **where** data comes from and **when**?
- Develop a **simple, generalisable model** that describes:
  - the **volume of content** posted at any given sample interval
  - the relative contribution of each of the 24 possible **time zones**
  - the expected **user behaviour** throughout a 24h period



# Data Generation Model

## Identifying the dominant frequencies

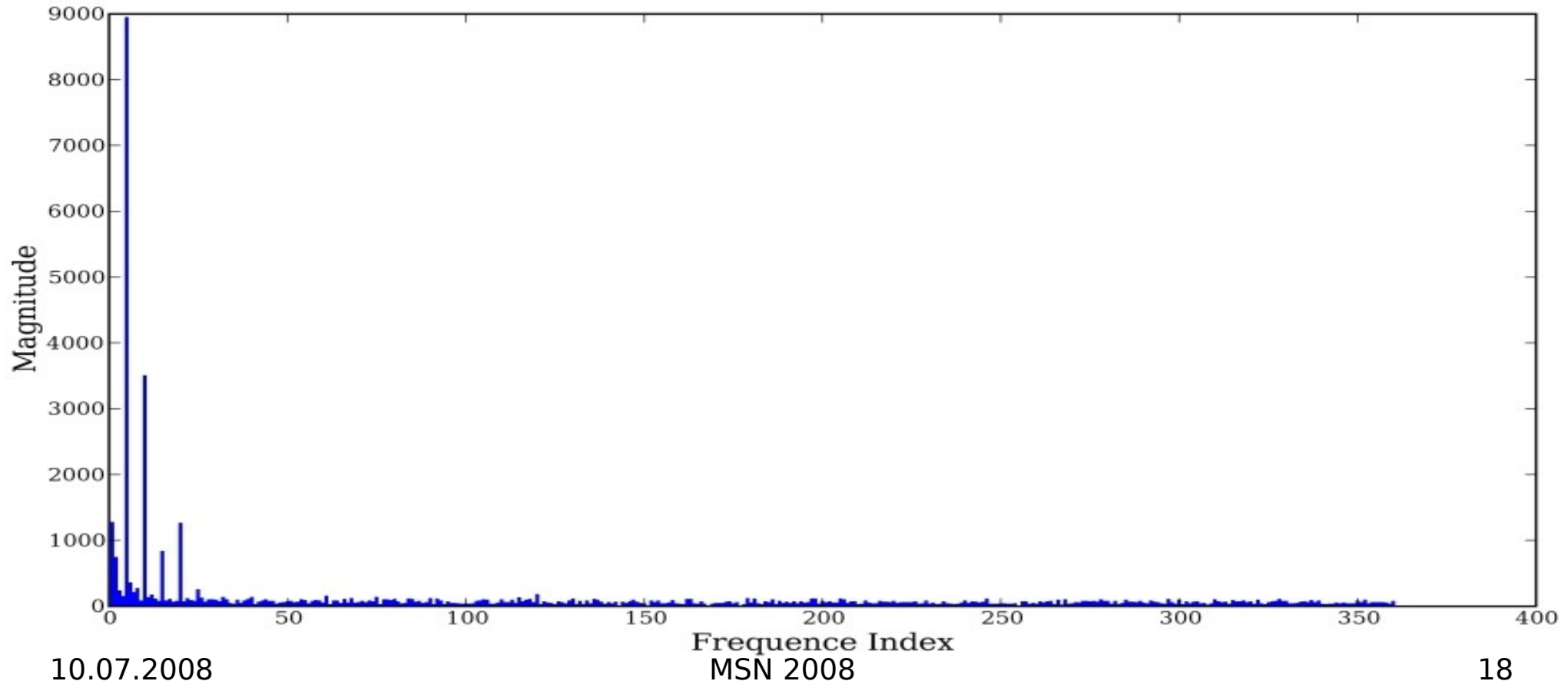
**Problem:** Unprocessed time series is noisy



# Data Generation Model

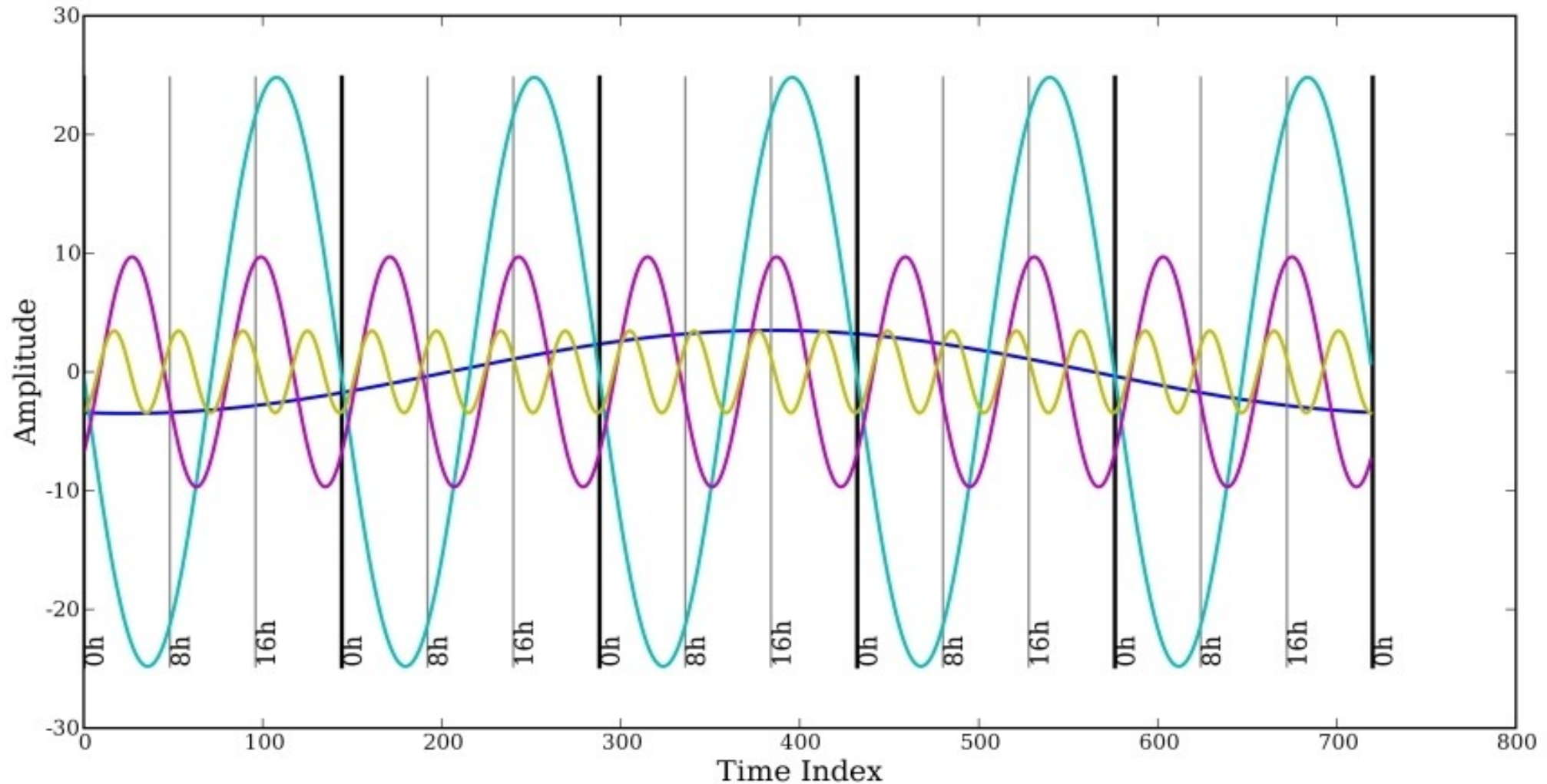
## Identifying the dominant frequencies

**Method:** Apply *Fourier Transformation* to identify the dominant frequencies.



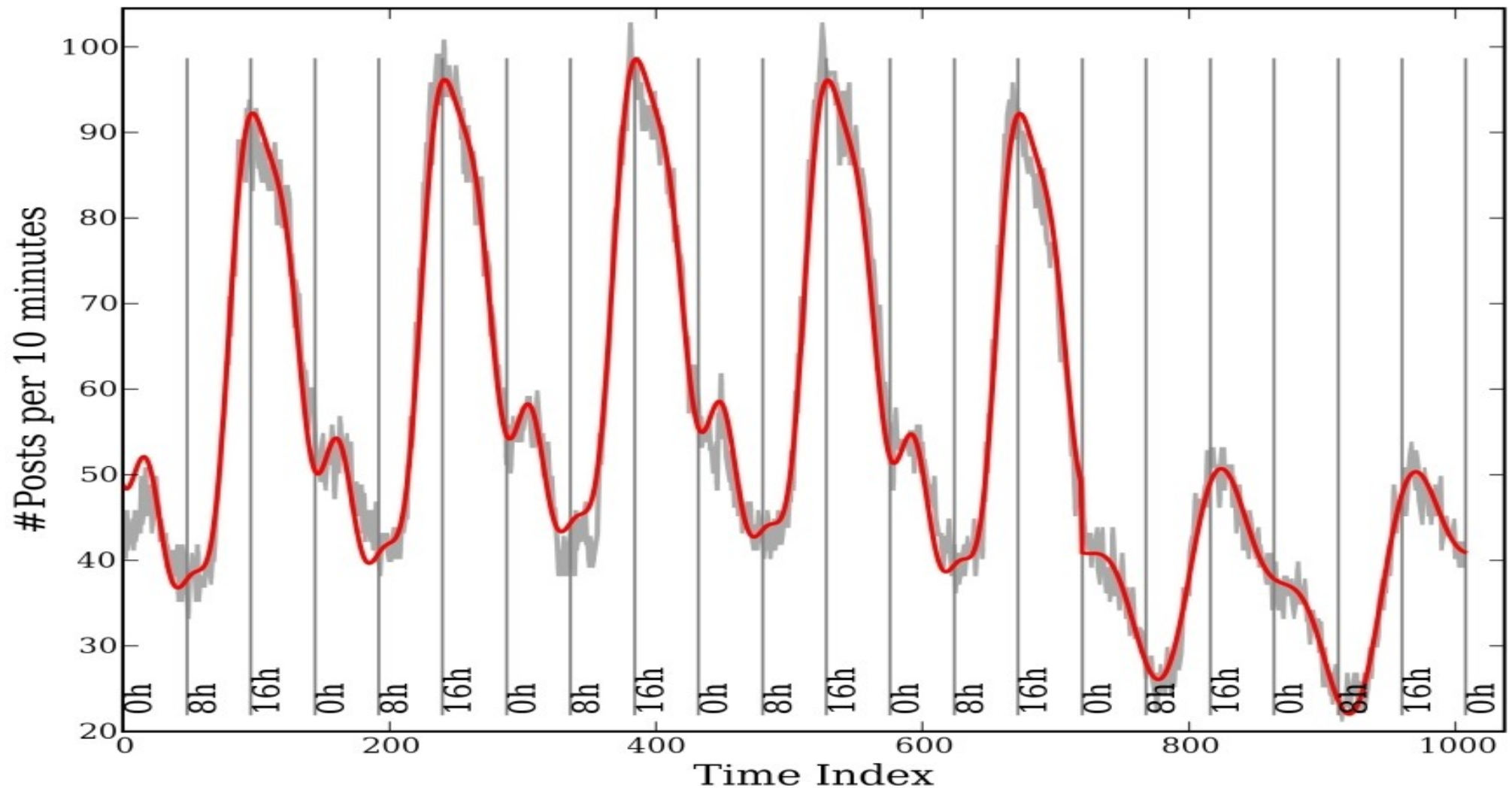
# Data Generation Model

## Identifying the dominant frequencies



# Data Generation Model

## Identifying the dominant frequencies



# Data Generation Model

## Step 2: time zone distribution

- **Problem:**

- Fourier gives us dominant frequencies, but no information from *where* the content was submitted.

- **Method:**

- Incorporate user *location information* into the Fourier model.

- **Assumptions:**

- Majority of users state correct location
- Users that do not reveal location are proportionally distributed in their geographical location

# Data Generation Model

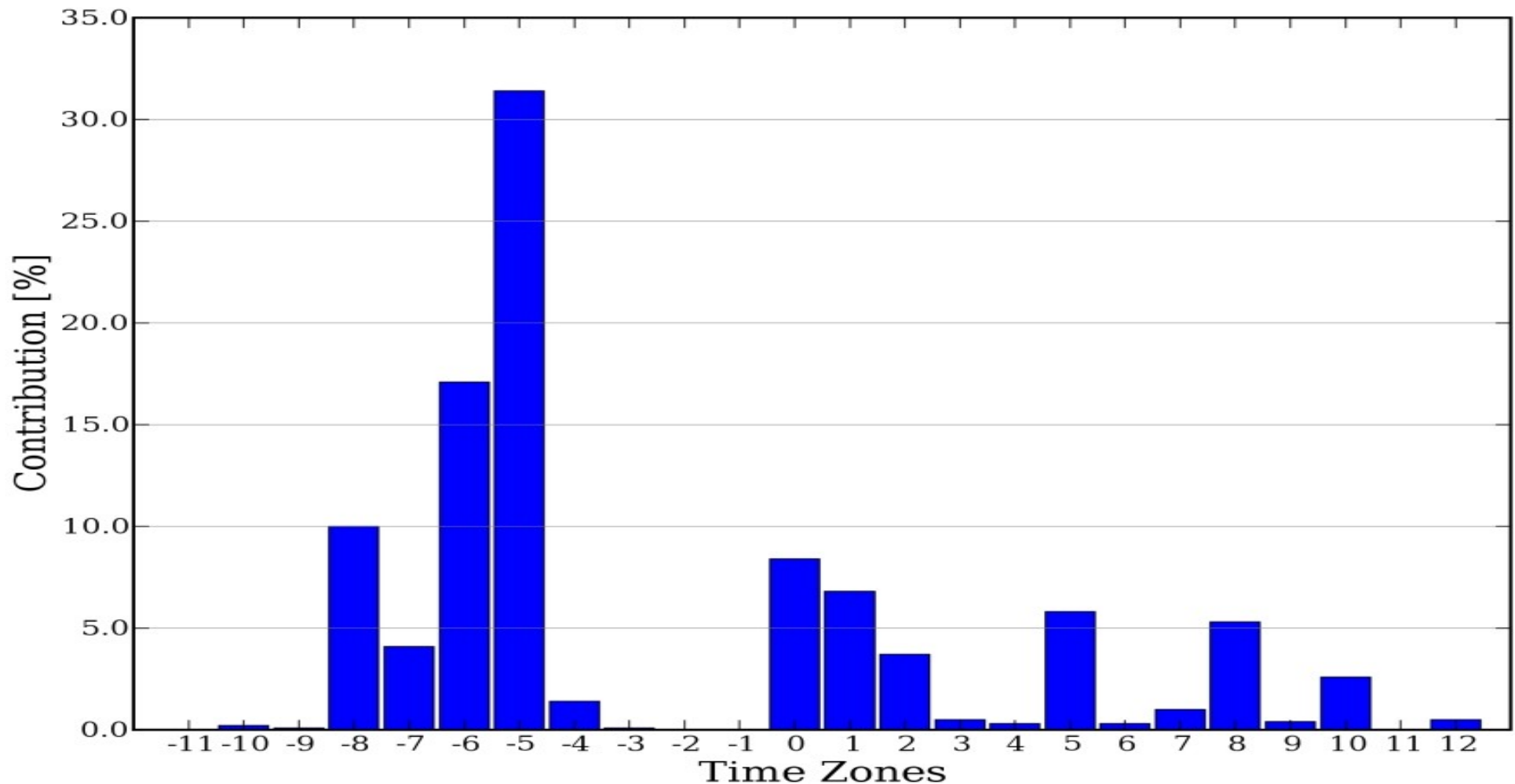
## Step 2: time zone distribution

Country	Time Zone (GMT)	Weight
United States	+10, -4 to -11	0.588
United Kingdom	0	0.075
Canada	-3 to -8	0.050
India	+5:30	0.036
Australia	+8 to +10:30	0.027
Germany	+1	0.013
France	+1	0.011
Italy	+1	0.010
China	+8	0.009
Brazil	-2 to -5	0.008

- **Problem:**  
Some countries have more than 1 time zone
- **Assumption:**  
User distribution is the same as popularity distribution within the zones

# Data Generation Model

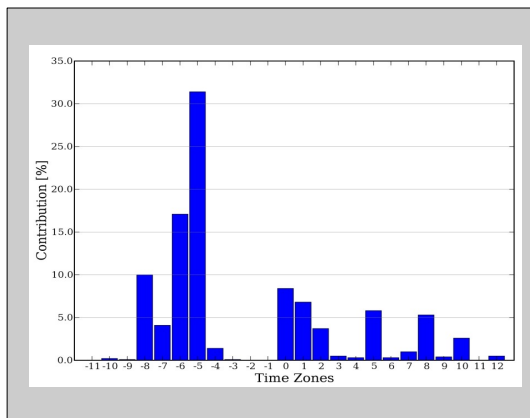
## Step 2: time zone distribution



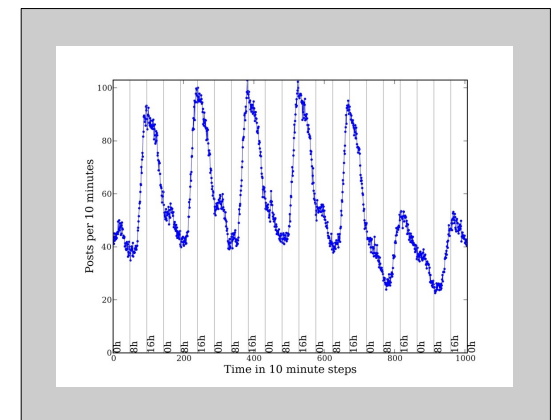
# Data Generation Model

## Step 3: expected user behaviour

- **Idea:**
  - Content volume per time interval is the sum of contribution of all time zones
- **Assumption:**
  - Users in different zones follow roughly the same usage pattern



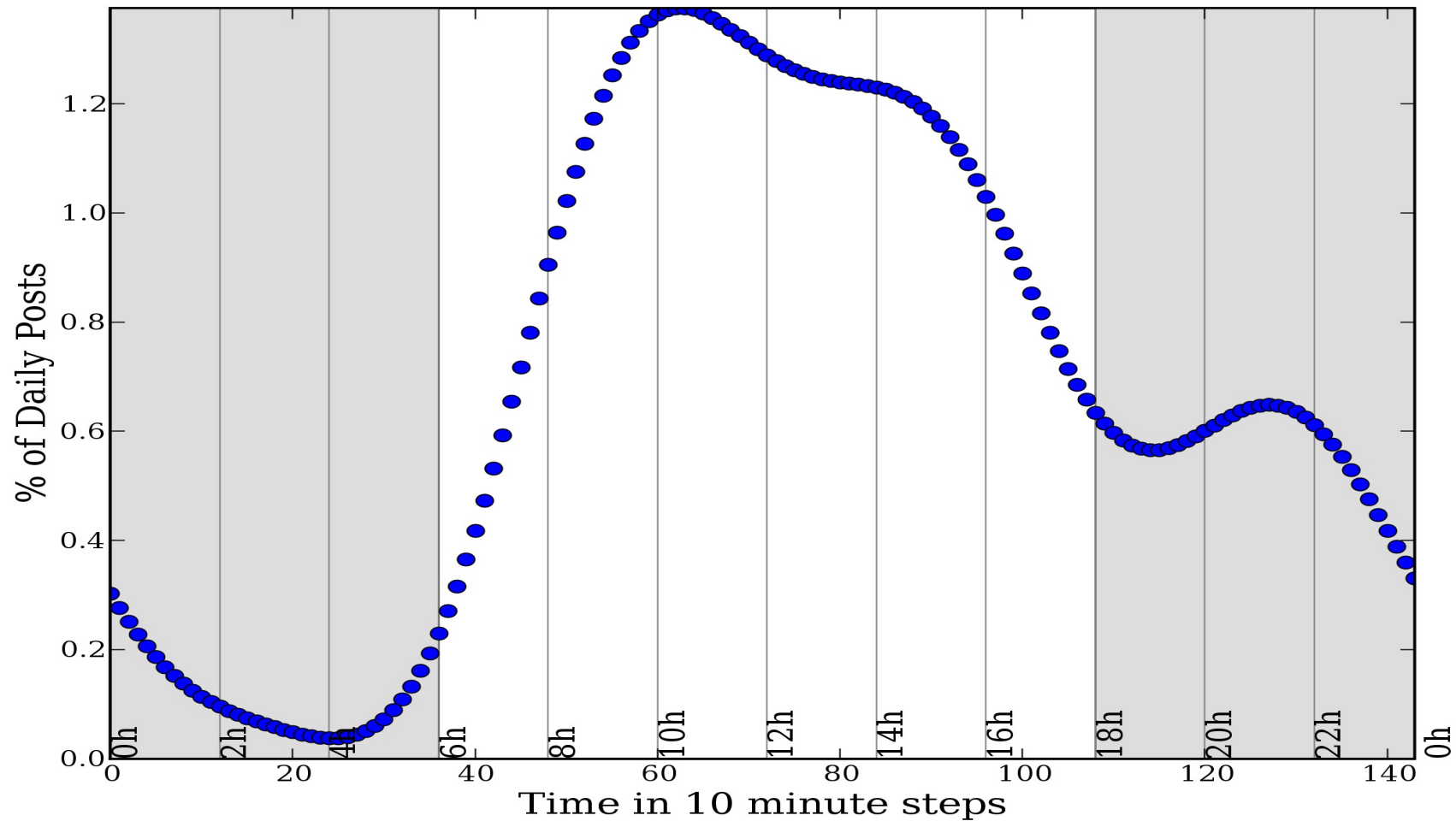
X ? =





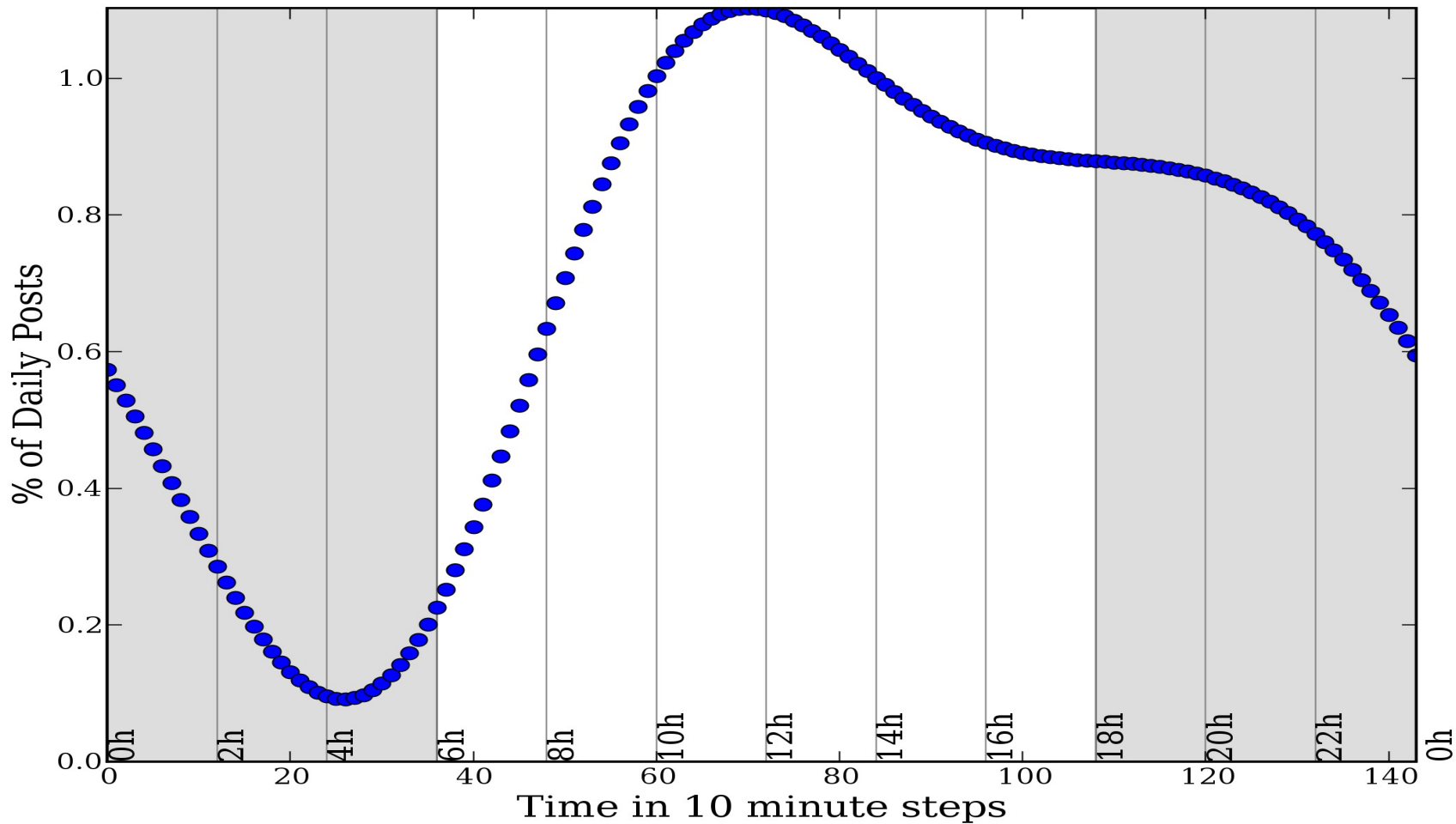
# Data Generation Model

## Step 3: expected user behaviour



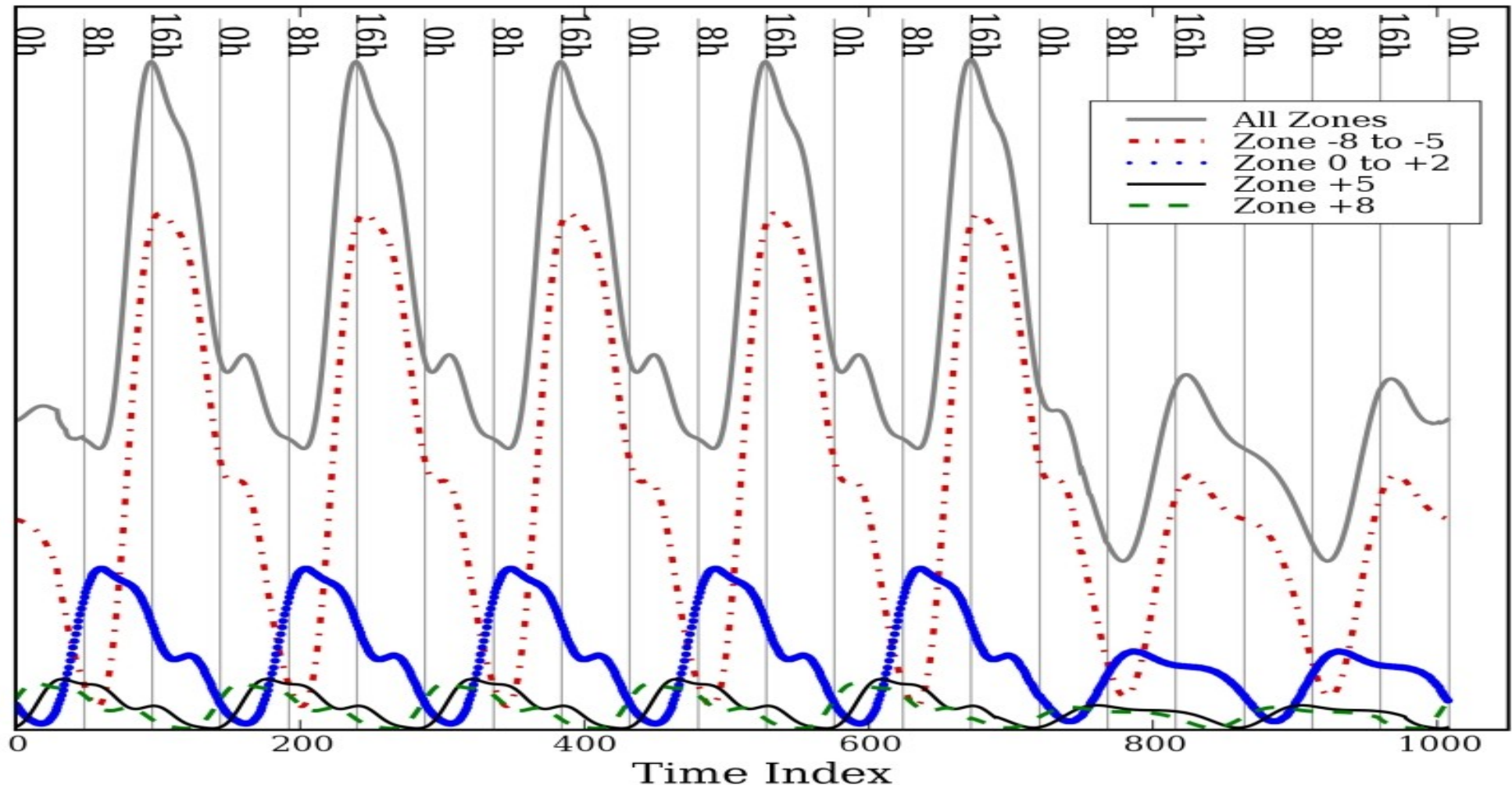
# Data Generation Model

## Step 3: expected user behaviour



# Data Generation Model

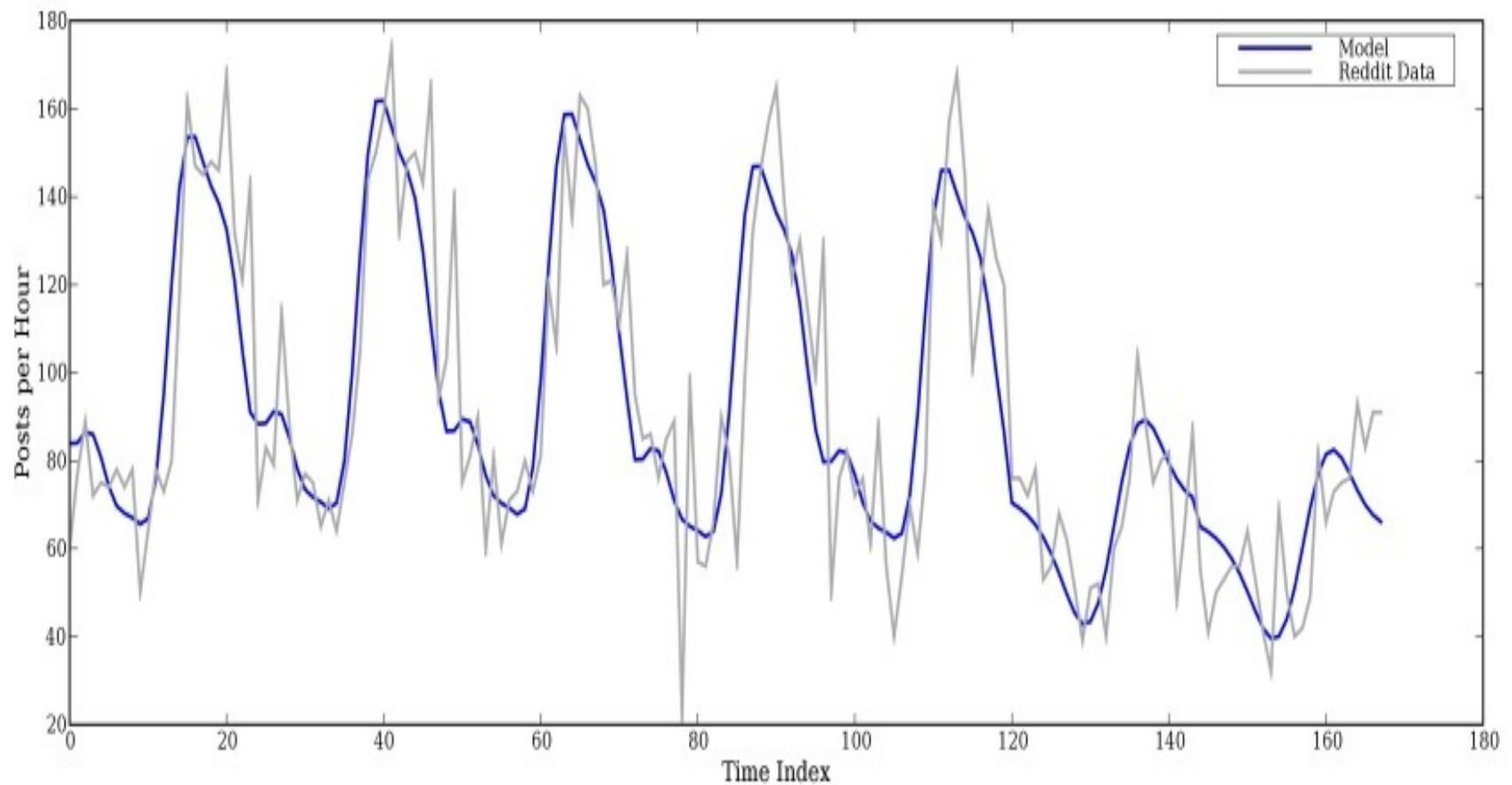
## Step 3: expected user behaviour



# Data Generation Model

## Model applied to reddit.com

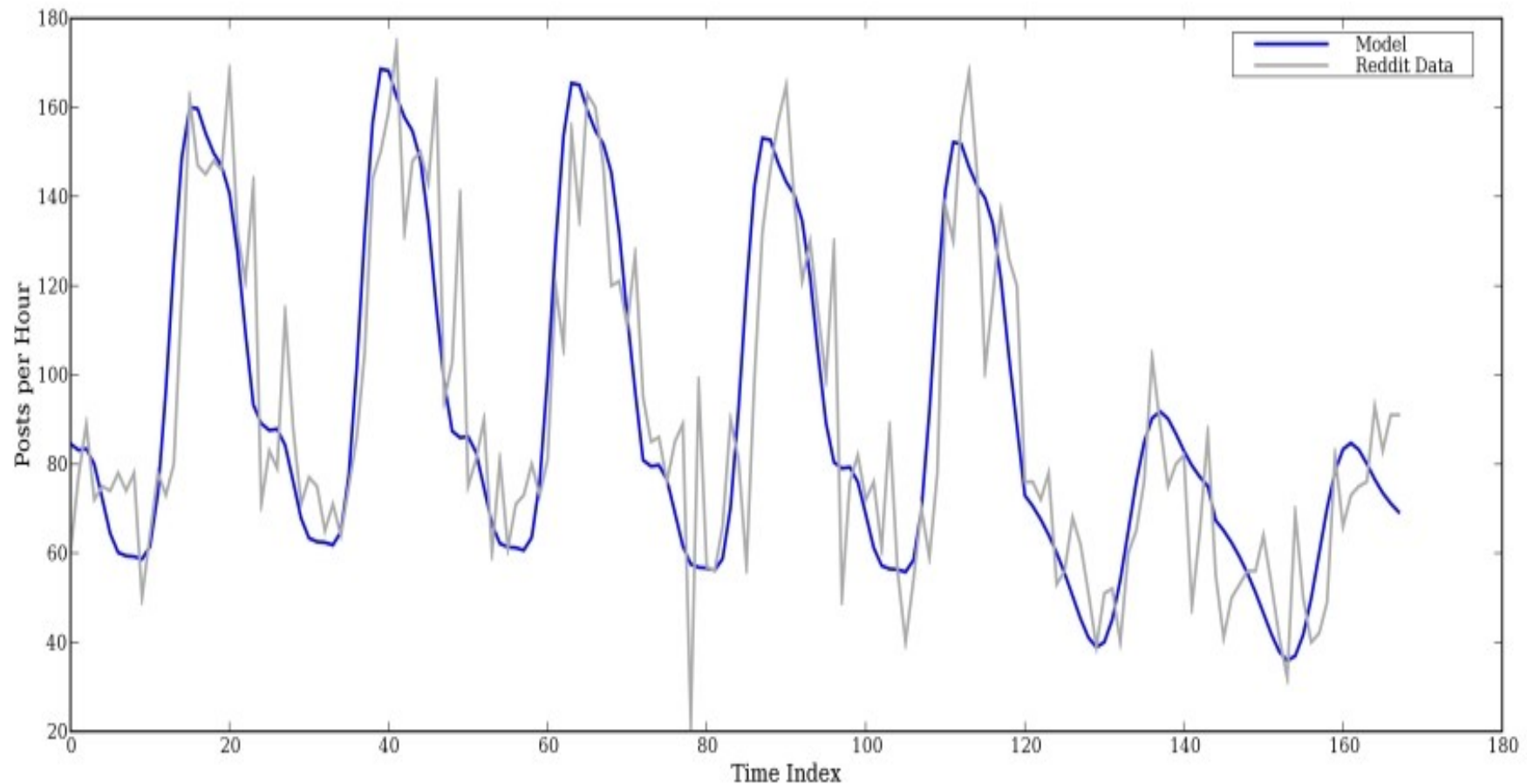
**initial fit:**



# Data Generation Model

## Model applied to reddit.com

**adapted weights:**



# Model Summary

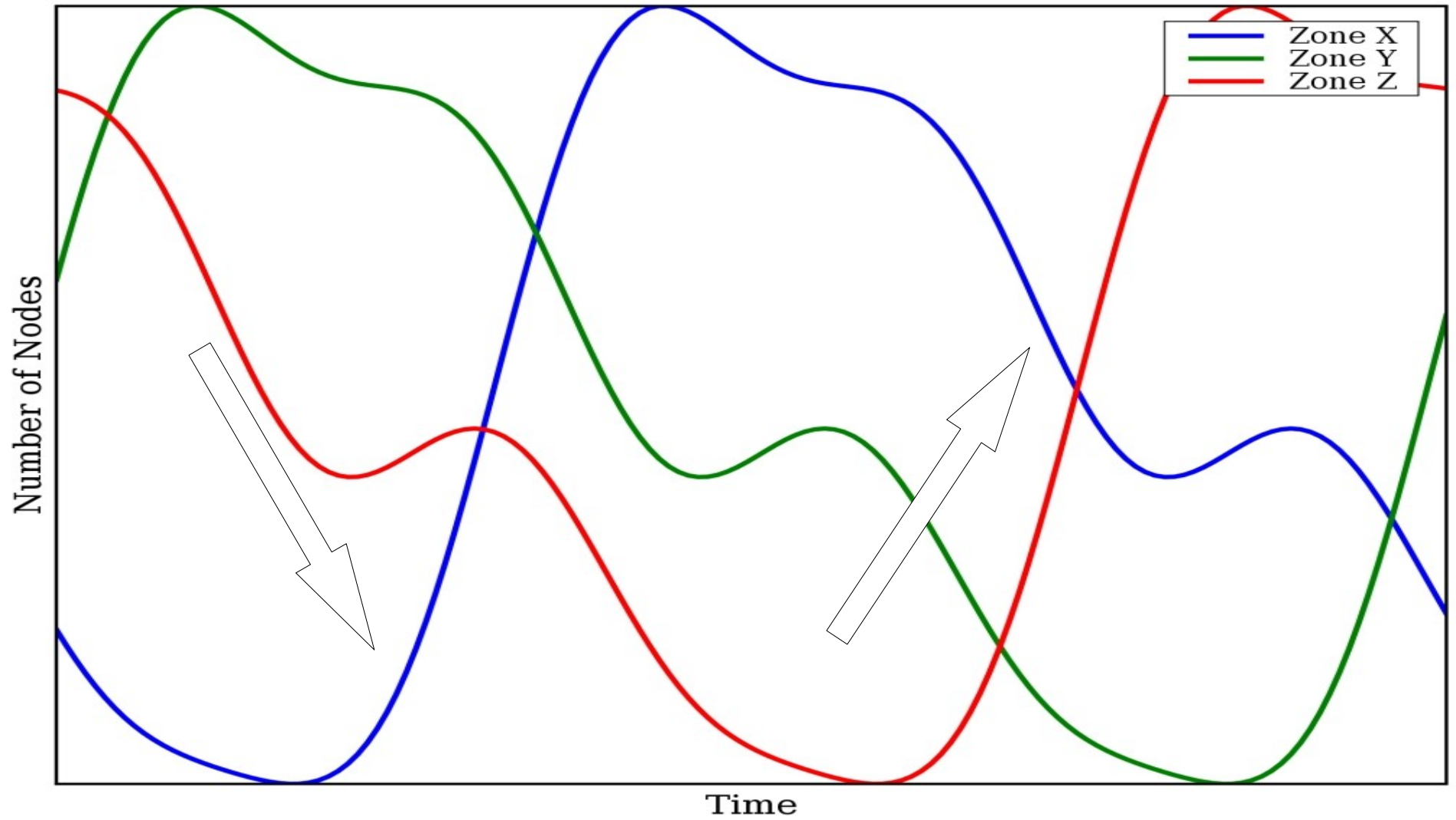
- Periodic pattern can be modelled with few dominant frequencies
- Time zone analysis reveals where content comes from
- Decomposed model describes user behaviour within a single time zone

# Design Implications

## Applying Geo-Temporal Information

- Energy-efficient load balancing
  - (Chen et al, NSDI 2008)
- Similar patterns exhibited in
  - Facebook (Golders et al, CT 2007)
  - MSN (Chen et al, NSDI 2008)
  - Gaming (Chambers et al, IMC 2005)
- Peer-to-Peer Churn / Content Distribution
  - neighbour selection / replication

# Example





# Future Work

- Comparing different node selection strategies when replicating data in distributed systems
- Can taking into account time zone information increase performance?
- Test other datasets
- How can time zone behaviour be learned in a distributed way?



# **The End**

# Thank you