EVOLUTION OF EXPERTS IN QUESTION ANSWERING COMMUNITIES

Aditya Pal, Shuo Chang, Joe Konstan
(University of Minnesota)

Motivation for Temporal Analysis

- Understanding user activity patterns
  - [Guo, 2009] analyzed hourly activity patterns to find that even though 80-20 contribution rule applies, yet top contributors' participation is much flatter than power law
- Question routing schemes
  - [Liu and Agichtein, 2011] showed that temporal activity patterns can be used effectively to tune question routing algorithms to ensure that a question gets answered in a timely manner

[KDD 2009] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. E. Zhao: Analyzing patterns of user content generation in online social networks.

[ECIR 2011] Q. Liu, and E. Agichtein: Modeling answerer behavior in collaborative question answering systems.

Research Questions

- How do experts evolve and influence community members ?
- What are the different evolutionary characteristics of experts ?
- Can we identify different kinds of experts ?
- Can we improve expert identification techniques by taking users' evolution into account?

Dataset Description

- StackOverflow data (August 2008 – September 2010)
  - ~ 1M question asked by 165K users
  - ~ 2.4M answers by 156K users

- Expert labeling
  - Selected users with more than 9 answers (29K users)
  - Marked top 10% with highest reputation score

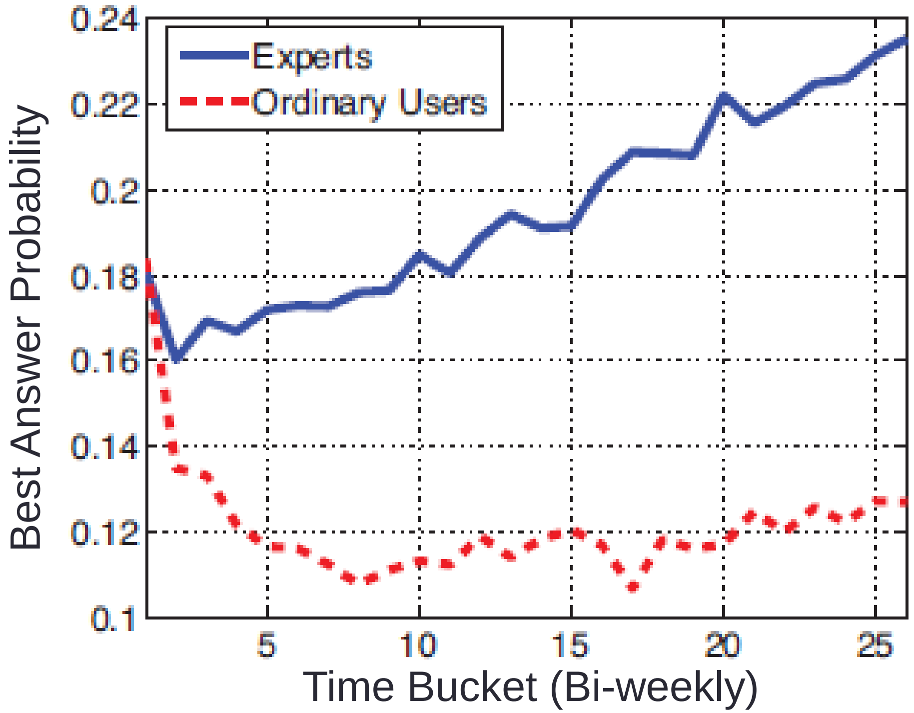
http://blog.stackoverflow.com/2010/09/creative-commons-data-dump-sept-10/

Data Preprocessing

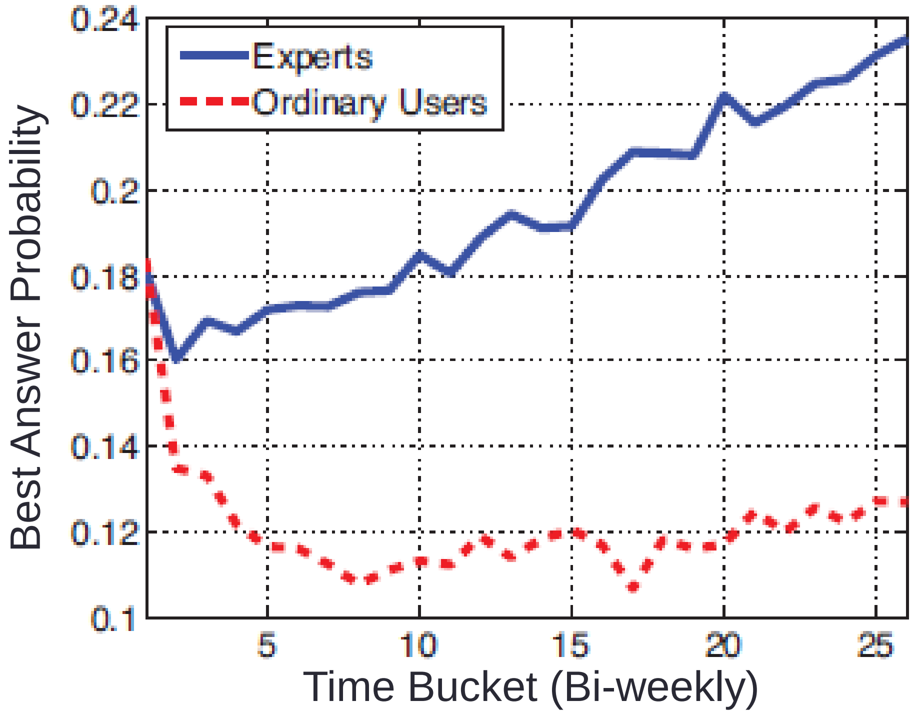
- Divide data into bi-weekly buckets
  - First bucket = time of earliest question
  - 70 bi-weekly buckets

- Relative time series
  - Pick first 26 buckets of activity for a user
  - Normalize based on activity of other users during the same time period

Influence on Question Askers

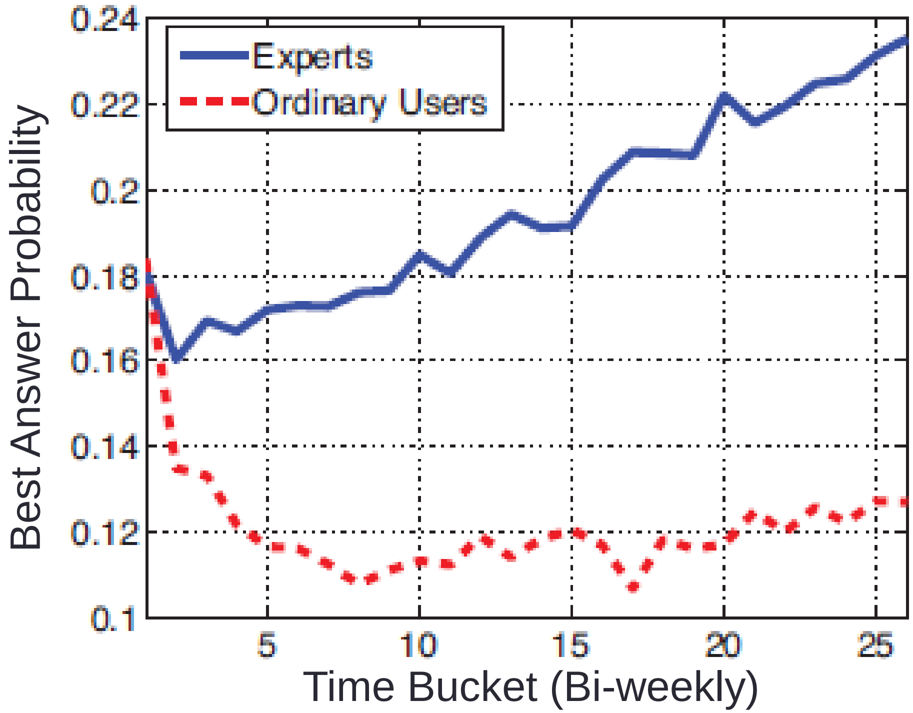


Influence on Question Askers



Askers are wary in selecting newcomers' answers as best

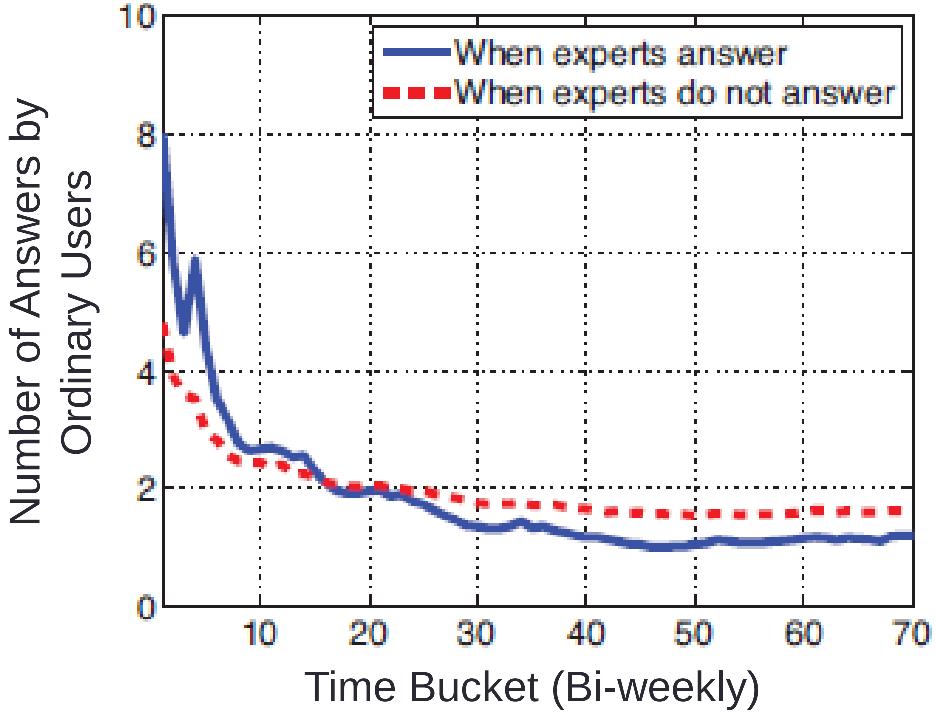
Influence on Question Askers



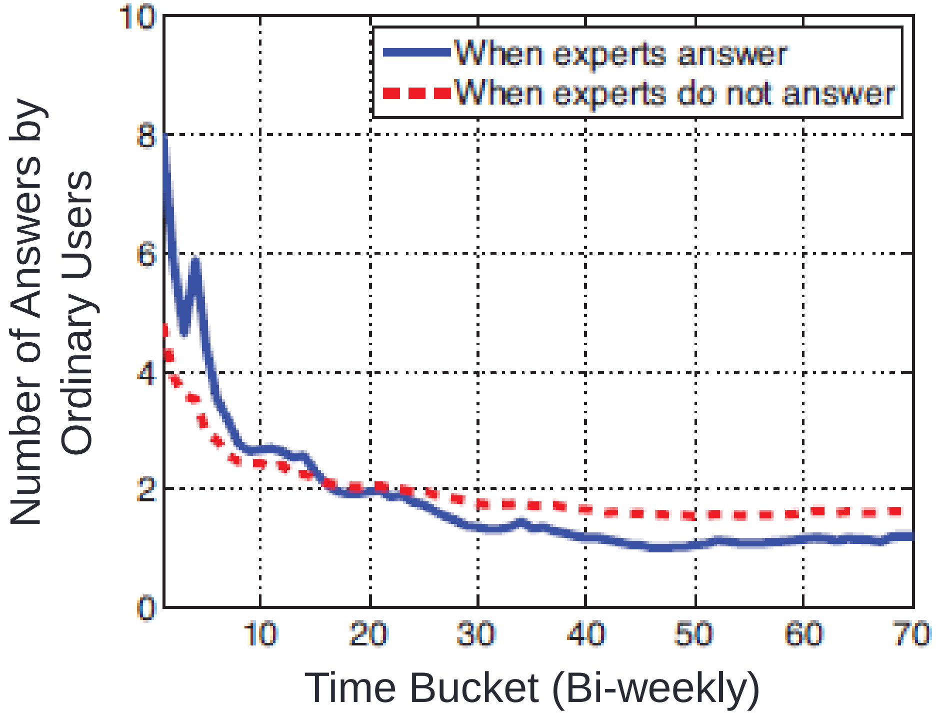
Askers are wary in selecting newcomers' answers as best

Experts get motivated as they get recognized for their work

Influence of Experts on Ordinary Users



Influence of Experts on Ordinary Users



Initially users participated vigorously on questions answered by experts

As experts became distinguishable, participation propensity decreased

Influence of Experts on Ordinary Users

- Prior work [Pal and Counts, ICWSM 2011]
  - Users get biased based on name value of experts

[ICWSM 2011] Pal and Counts: *What's in a @name? How name value biases judgment of microblog authors.*

Influence of Experts on Ordinary Users

- Prior work [Pal and Counts, ICWSM 2011]
  - Users get biased based on name value of experts

- Discussions in *meta*-StackOverflow
  - *Enormous contributions by experts demoralized them a bit*
  - *Its intimidating initially but with time one can adapt amongst experts*
  - *Its intimidating to answer a question asked by an expert*
  - *Merits and demerits of allowing easy questions to be answered by beginners*

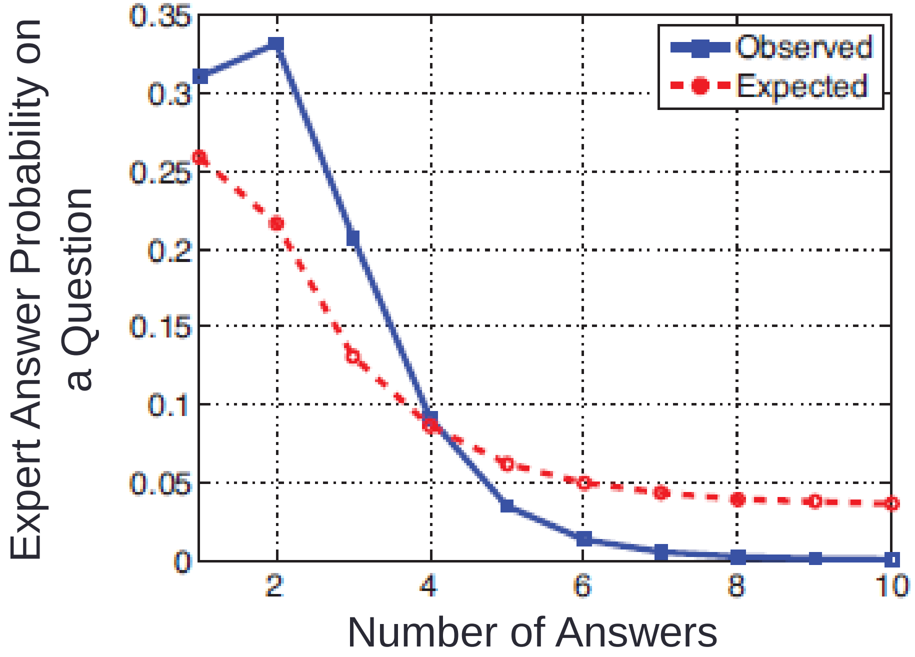
[ICWSM 2011] Pal and Counts: *What's in a @name? How name value biases judgment of microblog authors.*

Influence of Experts on Experts

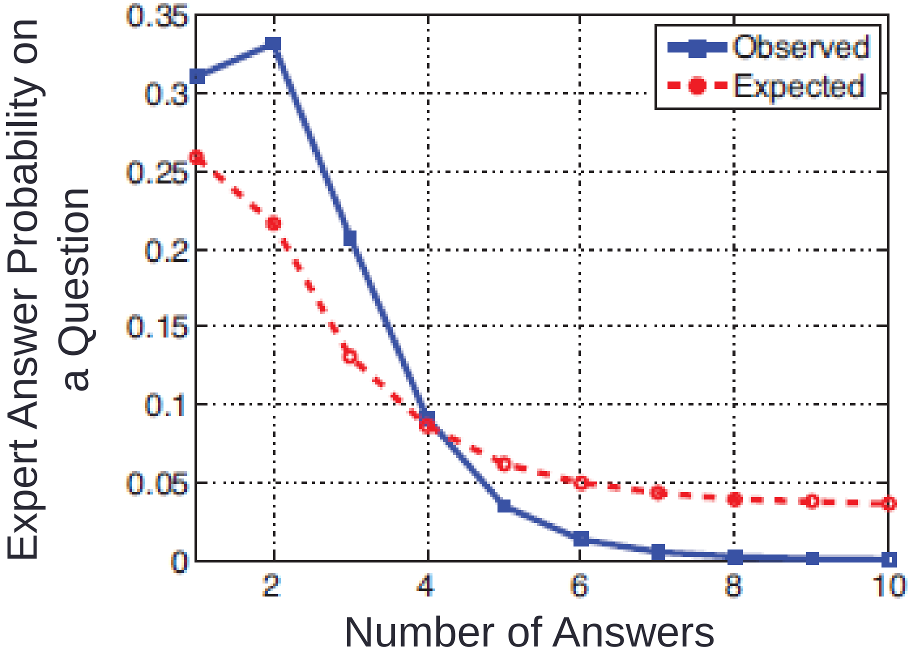
- p ~ probability of an expert answer (~0.4)
- n ~ number of answers to a question
- ne ~ number of expert answers to a question

$$\sim Binomial(n,p) = \frac{n!}{ne! \cdot (n-ne)!} p^{ne} (1-p)^{n-ne}$$

Influence of Experts on Experts

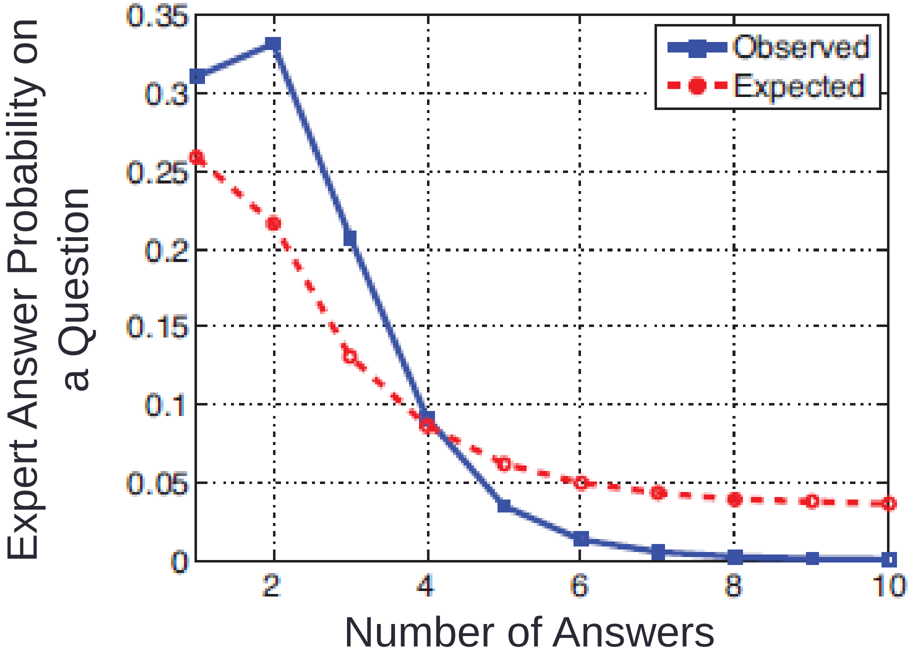


Influence of Experts on Experts



Experts are less likely to collectively collaborate to answer a question

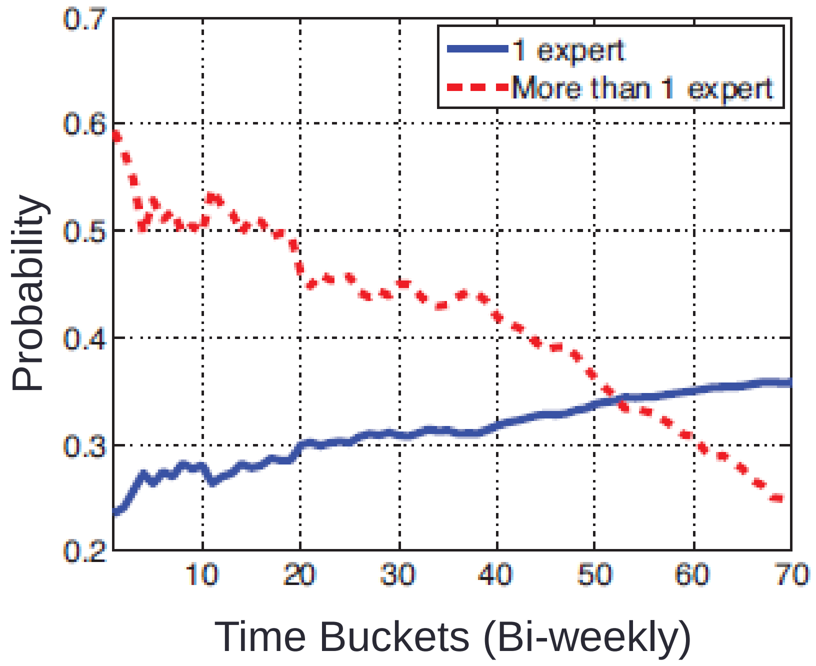
Influence of Experts on Experts



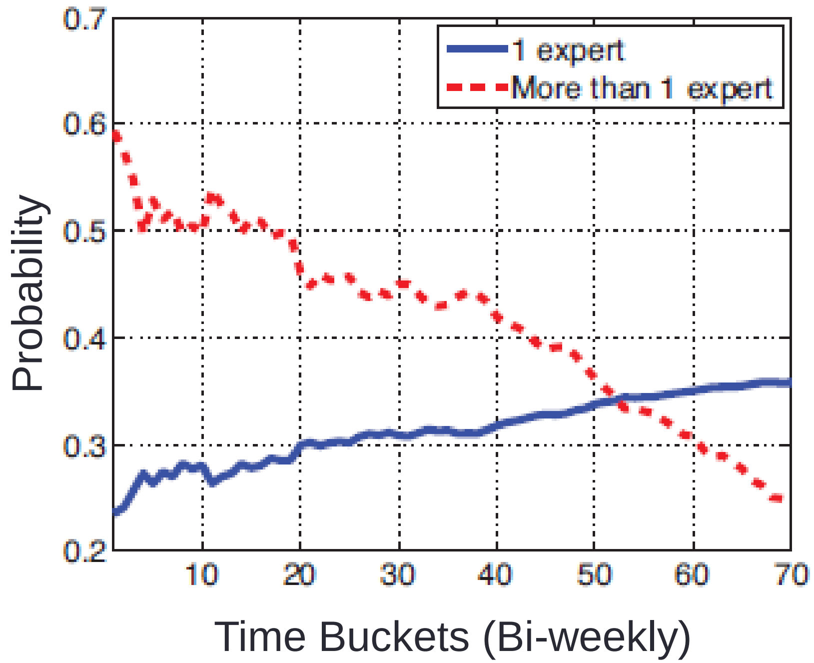
Experts are less likely to collectively collaborate to answer a question

Experts avoid each other, as they aim to have higher value/effort returns

Influence of Experts on Experts

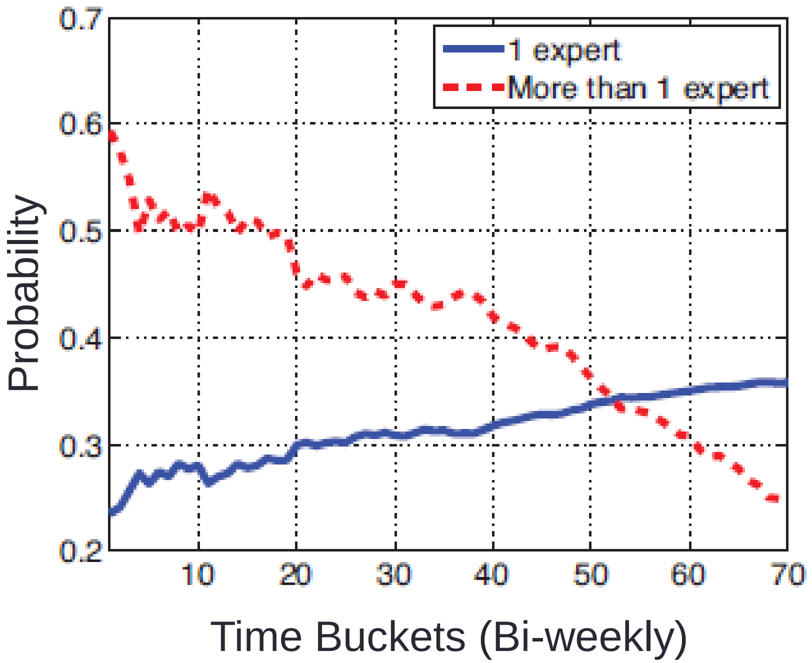


Influence of Experts on Experts



Initially, experts collaborated on more than 50% of the questions

Influence of Experts on Experts



Initially, experts collaborated on more than 50% of the questions

As experts became distinguishable, collaboration declined drastically

Expert Evolution

- Temporal Clustering Based on GMM

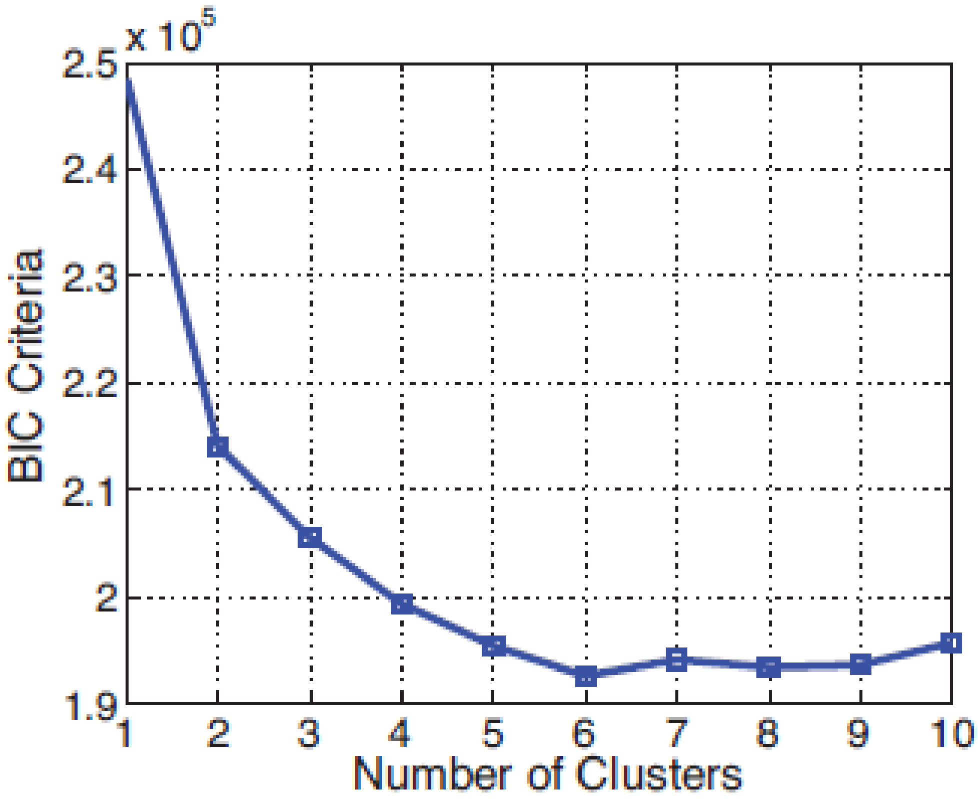
$$P(X|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_{ik} \cdot P(x_i|\theta_k) \quad \text{[i.i.d time series]}$$

$$P(x_i|\theta_k) \propto \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\,(x_i - \mu_k)^T\,\Sigma_k^{-1}\,(x_i - \mu_k)\right\}$$

- Bayesian Information Criteria

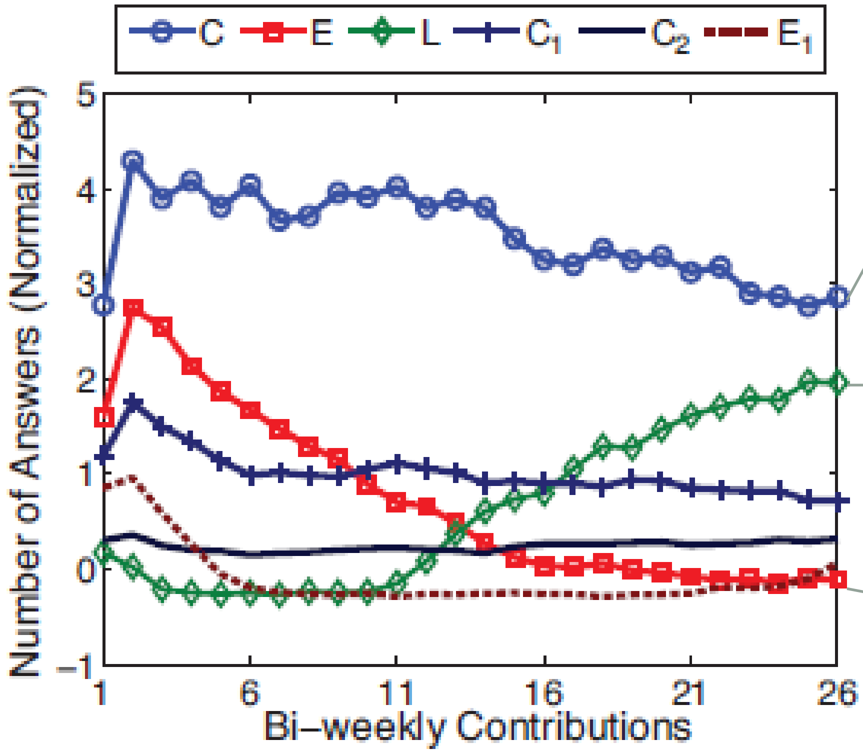
$$BIC(K) = -2 \cdot \ln\big(P(X|\theta)\big) + K \cdot \ln(N)$$

Number of Clusters



K=6 minimizes the BIC criteria

Expert Evolution Pattern



C = Consistently active

L = Initially passive, later active

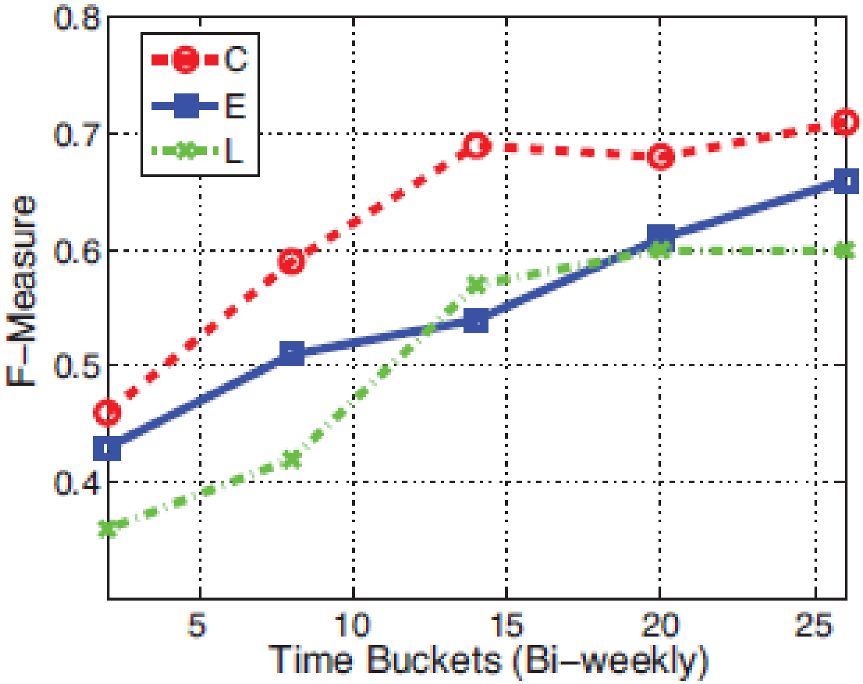
E = Initially active, later passive

For question routing, experts in C are valuable

For finding churners, experts in E are valuable

For nurturing and motivation, experts in L are valuable

Identifying Different Types of Experts



- SVM with 10-fold cross validation

$$\text{F-measure} = \frac{2 \cdot p \cdot r}{p + r}$$

Different types of experts can be found with 0.5 f-measure within 20-weeks of being in the community

Identifying Experts

	StackOverflow		Intuit	
	Static	Temporal	Static	Temporal
precision	90	**94**	70	**73**
recall	52	**67**	66	**71**
F-measure	66	**78**	68	**72**

Model based on temporal data outperform model based on static data

Summary

- Experts influence best answer selection of askers
- Ordinary users get intimated by experts
- Experts avoid other experts.
- Experts evolve with different patterns: C, E, L
- These experts can be found with satisfactory performance within 20 weeks
- Expert identification techniques can be improved by 5-15% by using their temporal data instead of static data

Take Aways

- Interface that anonymyze user profiles initially on a question can be beneficial. These profiles could be revealed after a lapse of time

- Different kinds of experts are useful for different objectives
  - C (consistently active) – question routing schemes
  - E (early active) – churn prediction
  - L (late active) – nurturing and fostering

- Expert identification methods can be improved by using temporal data

Thanks