On Crowdlearning: How do People Learn in the Wild?

Utkarsh Upadhyay

Joint work with Isabel Valera and Manuel Gomez Rodriguez

Max Planck Institute for Software Systems

What is crowdlearning?

Learning is becoming a crowd phenomenon.

What is crowdlearning?

Learning is becoming a crowd phenomenon.

Social learning

- Stack Overflow
- Quora
- Yahoo! Answers
- r/AskReddit, r/AskScience

What is crowdlearning?

Learning is becoming a crowd phenomenon.

Social learning

- Stack Overflow
- Quora
- Yahoo! Answers
- r/AskReddit, r/AskScience

Crowdlearning

When a crowd *learns* from knowledge *curated* and *contributed* by the crowd.

What is crowdlearning?

Learning is becoming a crowd phenomenon.

Social learning

- Stack Overflow
- Quora
- Yahoo! Answers
- r/AskReddit, r/AskScience

Crowdlearning

When a crowd *learns* from knowledge *curated* and *contributed* by the crowd.

Why should we care?

Crowdlearning

When a crowd *learns* from knowledge *curated* and *contributed* by the crowd.

- Understudied but useful and growing in importance
- *Just-in-time* learning instead of *just-in-case* learning
- Large amounts of data
- Learning is complex: insights may be transferable
  - Does it work?
  - *How much* knowledge is there?
  - Do known results hold?
  - Is it efficient/sustainable?
  - ...

A quick review of the process ...

Why should we care?

Crowdlearning

When a crowd *learns* from knowledge *curated* and *contributed* by the crowd.

- Understudied but useful and growing in importance
- *Just-in-time* learning instead of *just-in-case* learning
- Large amounts of data
- Learning is complex: insights may be transferable
  - Does it work?
  - *How much* knowledge is there?
  - Do known results hold?
  - Is it efficient/sustainable?
  - ...

A quick review of the process ...

Why should we care?

Crowdlearning

When a crowd *learns* from knowledge *curated* and *contributed* by the crowd.

- Understudied but useful and growing in importance
- *Just-in-time* learning instead of *just-in-case* learning
- Large amounts of data
- Learning is complex: insights may be transferable
  - Does it work?
  - *How much* knowledge is there?
  - Do known results hold?
  - Is it efficient/sustainable?
  - ...

A quick review of the process ...

Crowdlearning example: content

Knowledge Item

Smallest quantum of knowledge.

Crowdlearning example: content

Knowledge Item

Smallest quantum of knowledge.

Figure: An example knowledge item from Stack Overflow

Crowdlearning example: affordances

Contributions

Users can *contribute* to knowledge items.

Crowdlearning example: affordances

Contributions

Users can *contribute* to knowledge items.

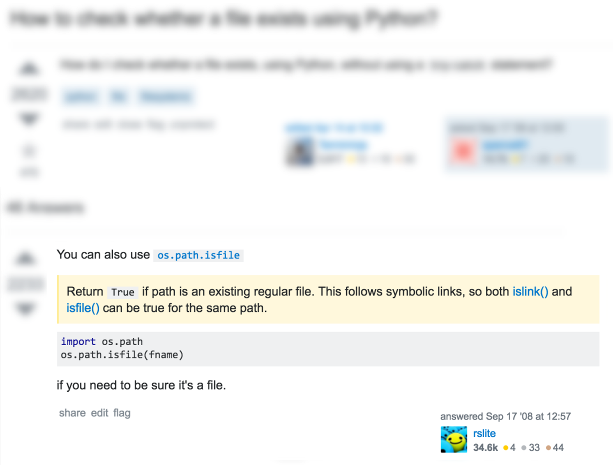


Figure: An answer contributed by a user.

Crowdlearning example: affordances

Learning events

Users can *learn* from knowledge items.

Crowdlearning example: affordances

Learning events

Users can *learn* from knowledge items.

Figure: A user learns from the knowledge item.

Crowdlearning example: process

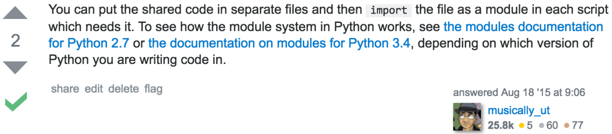


Figure: The same user later contributes knowledge on related topics

Crowdlearning example: final affordance

Assessment

The crowd *assess* the contribution made by other users.

Figure: Assessment of the contribution

Then the user reads more knowledge items, contributes, reads more, contributes, . . . all while increasing his expertise.

Crowdlearning example: final affordance

Assessment

The crowd *assess* the contribution made by other users.

Figure: Assessment of the contribution

Then the user reads more knowledge items, contributes, reads more, contributes, . . . all while increasing his expertise.

Key questions

Key Questions

- How does user expertise evolve?
- What is the true value of knowledge items?

Outline

Modeling Crowdlearning

Synthetic experiments

Results on real data
- Distribution of knowledge
- Types of learners
- Who learns where
- Who learns the most
- Who teaches better

Conclusion
- Limitations
- Future work

Data description

Learning events

An event which indicates reading of the knowledge items which *may* increase the expertise of a user.

$$l \quad := \quad (\underset{\substack{\uparrow \\ \text{user}}}{u}, \quad \underset{\substack{\uparrow \\ \text{time}}}{t}, \quad \overset{\substack{\text{knowledge item} \\ \downarrow}}{q})$$

e.g., upvoting an *answer* in a knowledge item.

Data description

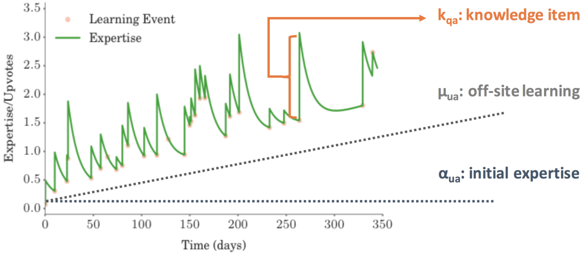
Contributing events

Contributions to a knowledge item which others can *assess*.

$$c \quad := \quad (\underset{\substack{\uparrow \\ \text{user}}}{u}, \quad \underset{\substack{\uparrow \\ \text{time}}}{t}, \quad \overset{\substack{\text{knowledge item} \\ \downarrow}}{q}, \quad \underset{\substack{\uparrow \\ \text{score}}}{s})$$

e.g., an answer and the upvotes it gets in the *first week*.

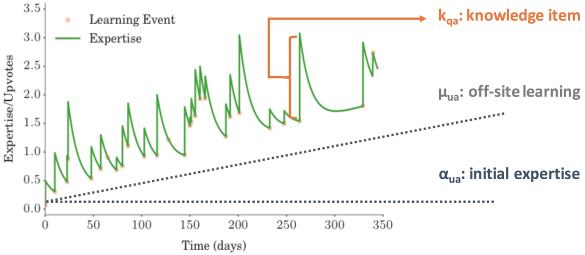
Crowdlearning Model



$$e^*_{ua}(t) := \overbrace{\alpha_{ua}}^{\text{initial expertise}} + \underbrace{\mu_{ua} \cdot t}_{\text{off-site learning}} + \overbrace{\sum_{i:q_i \in \mathcal{H}^l_u(t)} k_{q_i a} \cdot \underbrace{\kappa_\omega(t - t_i)}_{\text{forgetting}}}^{\text{on-site learning}}$$

Expertise of user u in topic a at time t.

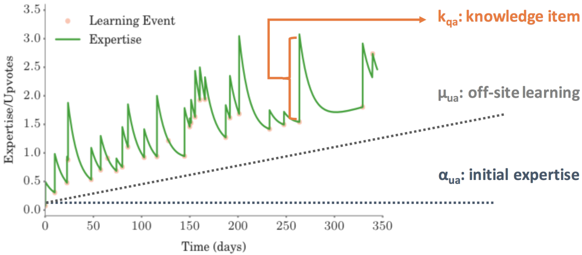
Crowdlearning Model



$$e^*_{ua}(t) := \overbrace{\alpha_{ua}}^{\text{initial expertise}} + \underbrace{\mu_{ua} \cdot t}_{\text{off-site learning}} + \overbrace{\sum_{i:q_i \in \mathcal{H}^l_u(t)} k_{q_i a} \cdot \underbrace{e^{-\omega(t-t_i)}}_{\text{forgetting}}}^{\text{on-site learning}}$$

Forgetting is modelled as an exponential.

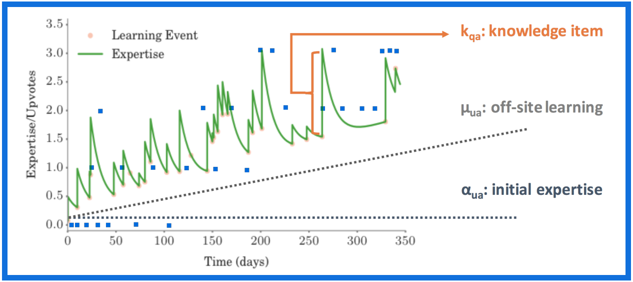
Crowdlearning Model



$$e^*_{ua}(t) := \overbrace{\alpha_{ua}}^{\text{initial expertise}} + \underbrace{\mu_{ua} \cdot t}_{\text{off-site learning}} + \overbrace{\sum_{i:q_i \in \mathcal{H}^l_u(t)} k_{q_i a} \cdot \underbrace{e^{-\omega(t-t_i)}}_{\text{forgetting}}}^{\text{on-site learning}}$$

- Expertise is *latent*.
- We can only observe assessments by others.

Crowdlearning Model



$$e^*_{ua}(t) := \overbrace{\alpha_{ua}}^{\text{initial expertise}} + \underbrace{\mu_{ua} \cdot t}_{\text{off-site learning}} + \overbrace{\sum_{i: q_i \in \mathcal{H}^l_u(t)} k_{q_i a} \cdot \underbrace{e^{-\omega(t - t_i)}}_{\text{forgetting}}}^{\text{on-site learning}}$$

$$p(score \mid \overbrace{\mathcal{A}_q}^{\text{Topics in } q}, \boldsymbol{e}^*_u(t)) \sim \text{Poisson}\left(\frac{\overbrace{\boldsymbol{w}_q^T}^{\text{Topics weights}} \cdot \boldsymbol{e}^*_u(t)}{\boldsymbol{w}_q^T \boldsymbol{1}} \right)$$

Parameter estimation

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{k}) = \sum_{\substack{(u,t,q,s) \\ \in \mathcal{H}^c(T)}} s \cdot \log \left(\frac{\boldsymbol{w}_q^T \boldsymbol{e}_u^*(t)}{\boldsymbol{w}_q^T \mathbf{1}} \right) - \frac{\boldsymbol{w}_q^T \boldsymbol{e}_u^*(t)}{\boldsymbol{w}_q^T \mathbf{1}}$$

$$\underset{\boldsymbol{\alpha} \geq 0, \boldsymbol{\mu} \geq 0, \boldsymbol{k} \geq 0}{\text{maximize}} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{k})$$

- Is jointly convex in $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$, and $\boldsymbol{k}$.
- Can be minimized using any convex optimization algorithm (we use L-BFGS-B).

Outline

Why synthetic experiments?

How many users/events will we need?

Synthetic experiments help us determine the subset of real data we can get accurate estimates for.

Two examples

- How many learning events do we need per knowledge item?
- How many contributions we need per user?

Why synthetic experiments?

How many users/events will we need?

Synthetic experiments help us determine the subset of real data we can get accurate estimates for.

Two examples

- How many learning events do we need per knowledge item?
- How many contributions we need per user?

Reliably estimate knowledge values

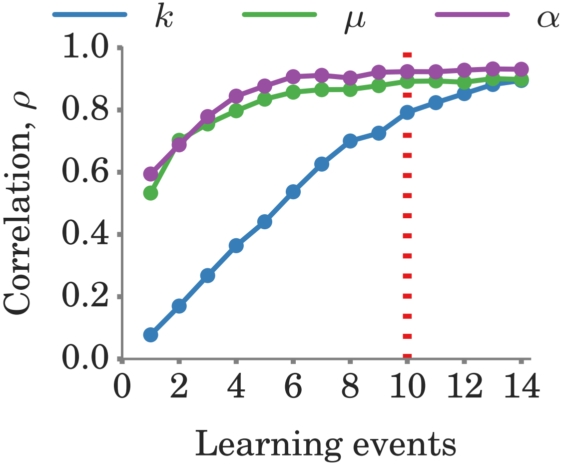


Figure: Need ≥10 learning events for a good estimation of knowledge values

Reliably estimate off-site learning rate

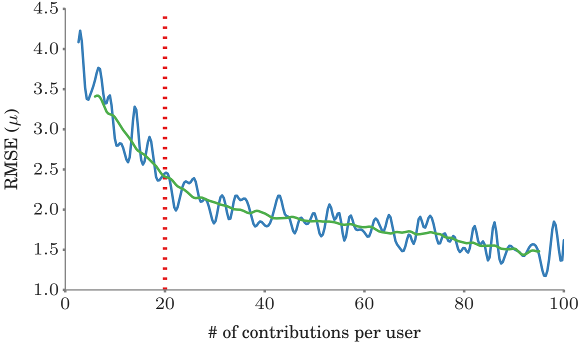


Figure: Need $\geq$20 contributions by each user for good estimation of μ

Outline

Data description

Mapping on Stack Overflow

Knowledge item	Question + all its answers
Learning event	Upvotes on answers
Contribution	An answer
Score	Upvotes received in 1^{st} week

We select top 10 tags (`java`, `c#`, `javascript`, `php`, `android`, `jquery`, `python`, `html`, `c++`, `mysql`)
After preprocessing, we have ∼**25k users** (with $\geq$ 20 contributions)

- who learn from ∼**66k knowledge items** (with $\geq$ 10 learning events)
  by means of **1.4m learning events**
- who contribute to 2.5m knowledge items
  by means of **3.8m contributions**.

Distribution of knowledge

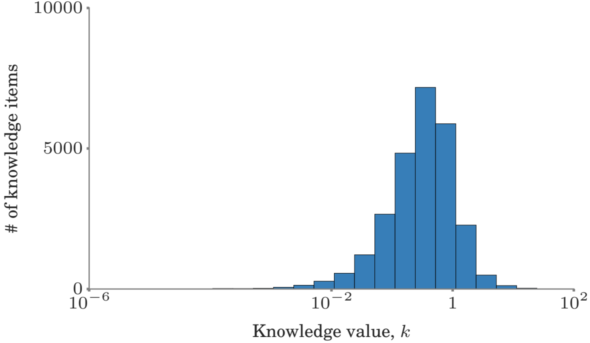
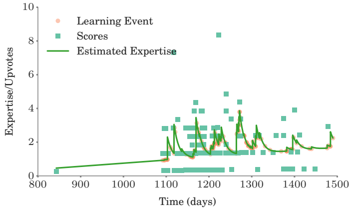


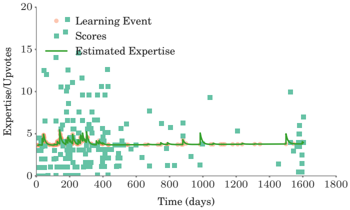
Figure: Distribution of knowledge values is log-normal

10% of knowledge items account for 75% of knowledge.

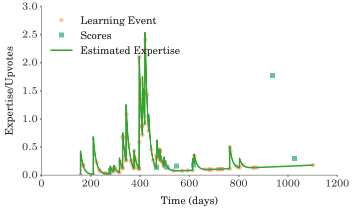
Types of learners



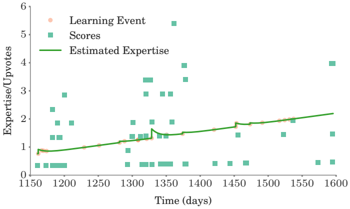
Avg. learner (Avg. knowledge / contribution: 0.005)



Expert: (Avg. knowledge / contribution: 0.034)



On-site learner (on-site learning: 55%)



Off-site learner (on-site learning: 0.4%)

Figure: Estimated learning trajectory for four characteristic Stack Overflow users

Where does learning happen

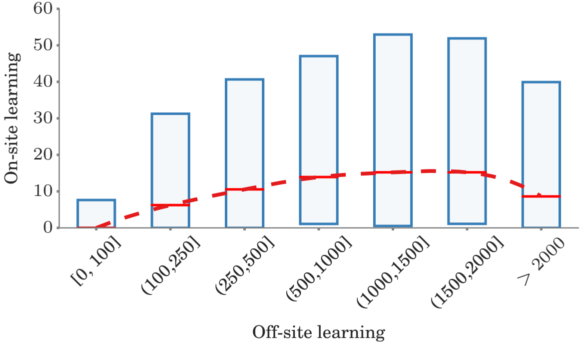
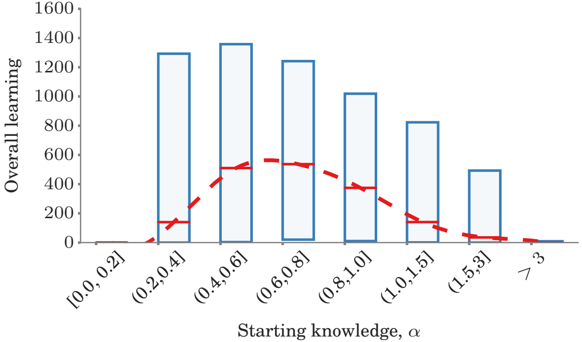


Figure: Users' on-site and off-site learning for c#

For $x \leq 2000$, users who achieve higher on-site learning also achieve higher off-site learning.
Over $x > 2000$, off-site learning becomes more dominant.

Who learns the most



- Newbies and experts increase their knowledge the least.
- Users in the middle of the range tend to increase it the most.[1]

[1] Leibowitz 2010

Who teaches better

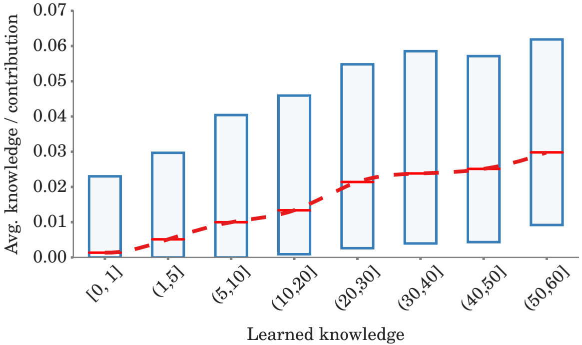


Figure: Avg. knowledge per contribution vs. learned knowledge

The users that learn more knowledge are also more proficient at producing high knowledge contributions.

By learning your will teach, by teaching you will learn.
– Latin proverb.

Who teaches better

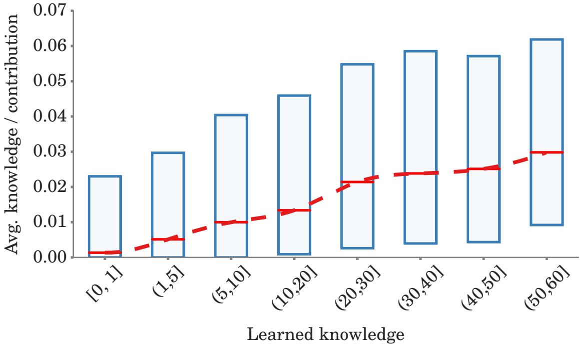


Figure: Avg. knowledge per contribution vs. learned knowledge

The users that learn more knowledge are also more proficient at producing high knowledge contributions.

By learning your will teach, by teaching you will learn.
– Latin proverb.

Outline

Limitations

Number of parameters

Large number of parameters require large amount of data.

Mapping learning events and assessments

It may be difficult to map site-events to learning events and to assessments of expertise.

- In Wikipedia, it may be difficult to find learning events.
- In Reddit, it may be difficult to determine topics for expertise tracking.

Limitations

Number of parameters

Large number of parameters require large amount of data.

Mapping learning events and assessments

It may be difficult to map site-events to learning events and to assessments of expertise.

- In Wikipedia, it may be difficult to find learning events.
- In Reddit, it may be difficult to determine topics for expertise tracking.

Conclusion

A model of crowdlearning

An expressive model which can:

- capture evolution of expertise,
- uncover true value of knowledge items, and
- scale up to web-sized datasets.

Future work

- Modeling overlaps in knowledge items
- Other crowdlearning networks (*e.g.*, citation graph)
- Merging data from other sources, *e.g.*, MOOCs

Questions?

Utkarsh Upadhyay
utkarshu@mpi-sws.org

Conclusion

A model of crowdlearning

An expressive model which can:

- capture evolution of expertise,
- uncover true value of knowledge items, and
- scale up to web-sized datasets.

Future work

- Modeling overlaps in knowledge items
- Other crowdlearning networks (*e.g.*, citation graph)
- Merging data from other sources, *e.g.*, MOOCs

Questions?

Utkarsh Upadhyay
`utkarshu@mpi-sws.org`

Overall learning

Given user u, we define the following.

On-site learning

The total expertise gathered by reading the knowledge items:

$$\sum_{a \in \mathcal{A}} \sum_{q \in \mathcal{H}_u^l(T)} \int k_{qa} \kappa_\omega(t)\, dt$$

Off-site learning

The expertise gathered outside Stack Overflow:

$$\sum_{a \in \mathcal{A}} \int \mu_{ua} t \, dt$$

Overall learning

Sum of on-site and off-site learning.

Changing forgetting rate

Half-life of knowledge

Time it takes to forget 50% of the knowledge from an item.

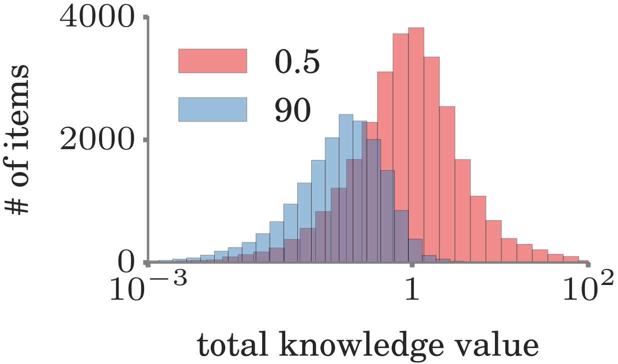
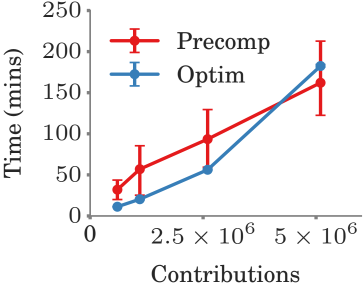
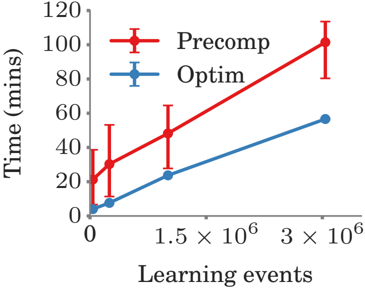


Figure: Fewer knowledge items have knowledge which lasts for longer periods of time

Scalability



Contributing events



Leraning events

Figure: Scalability

Evaluation

Score difference	# of pairs	Off-site only	Our model
≥ 1.0	31,639	52.5%	**61.9%**
≥ 2.0	19,253	52.9%	**64.8%**
≥ 3.0	10,804	53.2%	**67.0%**
≥ 4.0	5,910	53.7%	**70.7%**
≥ 5.0	3,250	55.0%	**71.6%**
≥ 6.0	1,935	56.0%	**73.3%**
≥ 7.0	1,159	56.8%	**73.8%**

Table: Effect of ignoring the knowledge variables on the accuracy of our model in predicting relative quality of competing answers. As the difference between the scores obtained by the answers increases, it should become easier to correctly identify the differences. This effect is more pronounced in our model than in model which only models off-site learning.

Evaluation

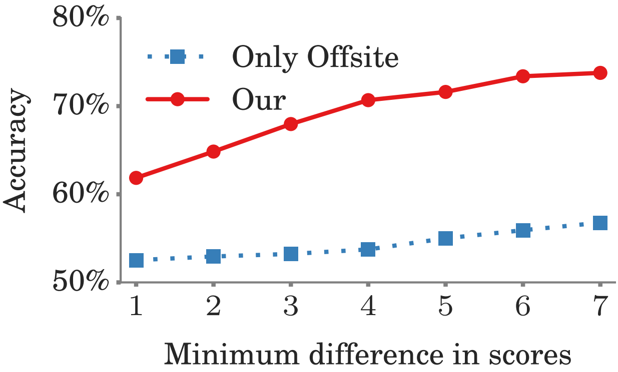


Figure: The performance of our model against the model which only models off-site learning for users.