

CiteRank

A Google-Inspired Ranking
Algorithm for Citation Networks

Dylan Walker

Brookhaven National Lab
Stony Brook University



Summary

- The Problem: Ranking Citation Networks
- Why the old way is bad
- How we can learn from Google
- CiteRank Model
- Performance on Real Citation Networks
- Optimal Parameters
- Why it works – physical interpretation

The old way of ranking publications

- Current method of ranking citation networks:

$k_{in} \sim$ the number of citations received

- But this is unfair:
 - New papers have not been around long enough to accrue citations
- All citations are not equal
 - A new citation should count more than an old one
 - Citations from popular papers should count more
- Google PageRank does this...

Google Predicts Traffic

- Why is Google's PageRank so successful?
 - How do we know it *is* successful?
 - PageRank is a model of traffic: The PageRank of a page can be interpreted as the predicted traffic for that page.
- 10^{10} heads are better than 1:
 - An ensemble of random surfers walk on the network.
 - Predictions of traffic to a given site are determined from the average visitation.
- Random surfers aren't smart, but the network is.
 - Walking on a network accounts for the self-consistence of popularity.
- So why can't we use Google on citation network?

Google and Citation Networks

- Citation networks are fundamentally different from the web
- Citation networks are acyclic and have an intrinsic time-arrow:
 - The links on a webpage can be updated at any moment. It is their own responsibility to maintain relevancy.
 - The citations in a publication remain fixed.
- What does this mean for ranking?
 - Given enough time, random researchers (surfers) would pile-up at the old edge of the network.
 - Aging effects cannot be ignored.
- Can we still model traffic on Citation Networks with random researchers?

The CiteRank Model of Traffic

- The CiteRank prediction of traffic has two parameters: $T_i(\alpha, \tau_d)$
- With a fixed probability, each researcher will follow a citation to an adjacent publication
 - Probability to follow a link $\sim (1 - \alpha)$
- Distribute random researchers on a citation network according to an initial distribution:
 - $\rho(\tau_d)$, where $\tau_d \sim$ characteristic decay time
- The CiteRank algorithm is given by:

$$\vec{T} = I \cdot \vec{\rho} + (1 - \alpha)W \cdot \vec{\rho} + (1 - \alpha)^2 W^2 \cdot \vec{\rho} + \dots = \frac{\vec{\rho}}{I - (1 - \alpha)\hat{W}}$$

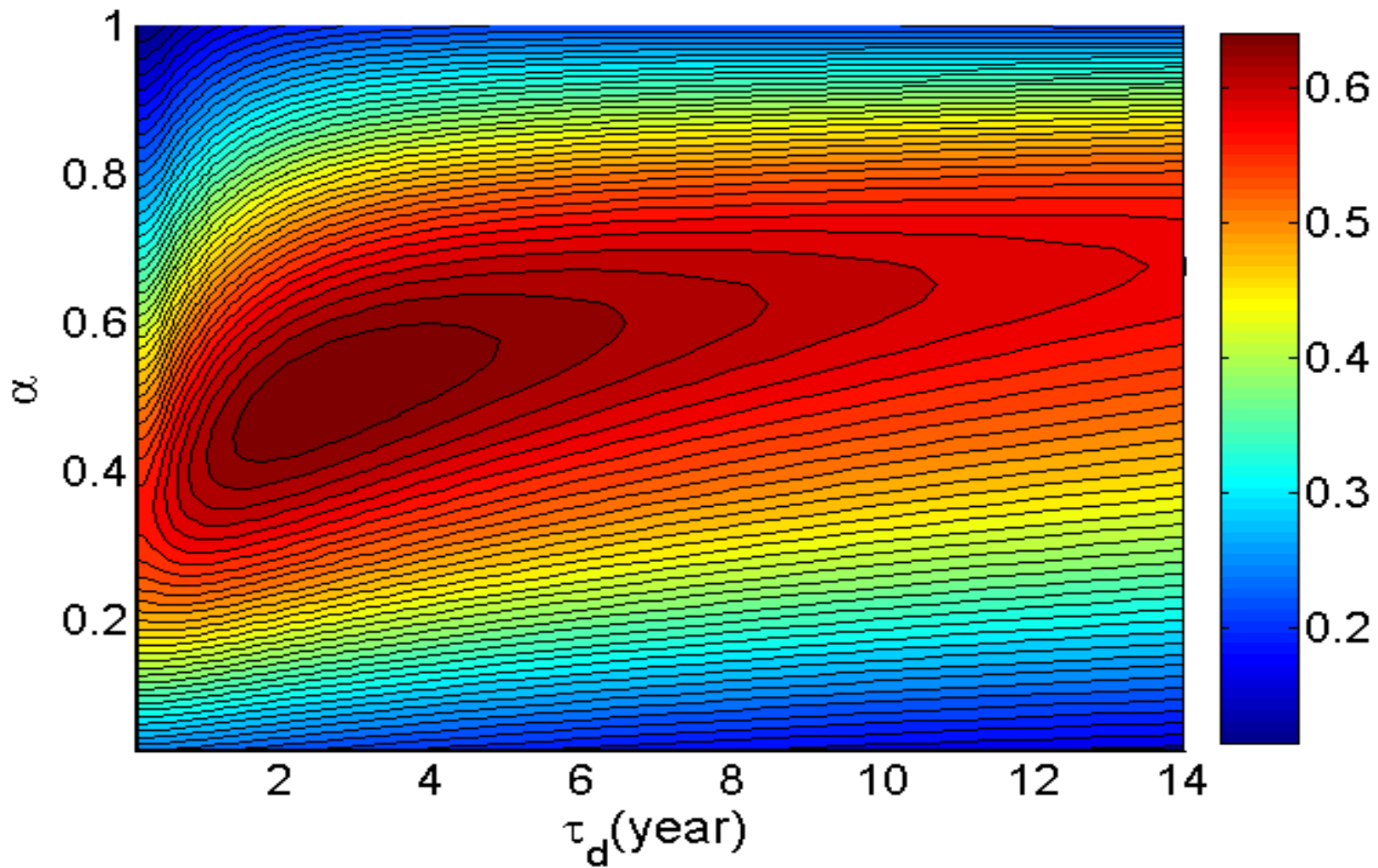
Two Real Citation Networks

- To select the best parameters and see if CiteRank is a viable ranking scheme, we evaluate two real citation networks:
- High Energy Physics Theory ArXiv (hep-th):
 - A snapshot of the high energy physics theory area of arxiv.org from April 2003 (citations ranging from 1992-2003)
 - 2800 papers ; 350,000 citations
 - no form of peer review
- Physical Review (physrev):
 - Citation data from all Physical Review journals (citations ranging from 1913-2005)
 - 380,000 papers ; 3,100,000 citations

CiteRank Optimal Parameters

- The CiteRank predicts traffic
- Ideally, we would like to select parameters that best correlate T_i with real traffic, T_i^{real} .
- However, traffic data is not readily available.
- Can estimate T_i^{real} with the recently accrued citations, Dk_i .
- Relationship between T_i^{real} and Dk_i is unclear:
 - Assume linearity and test the correlation over range of the model parameters.

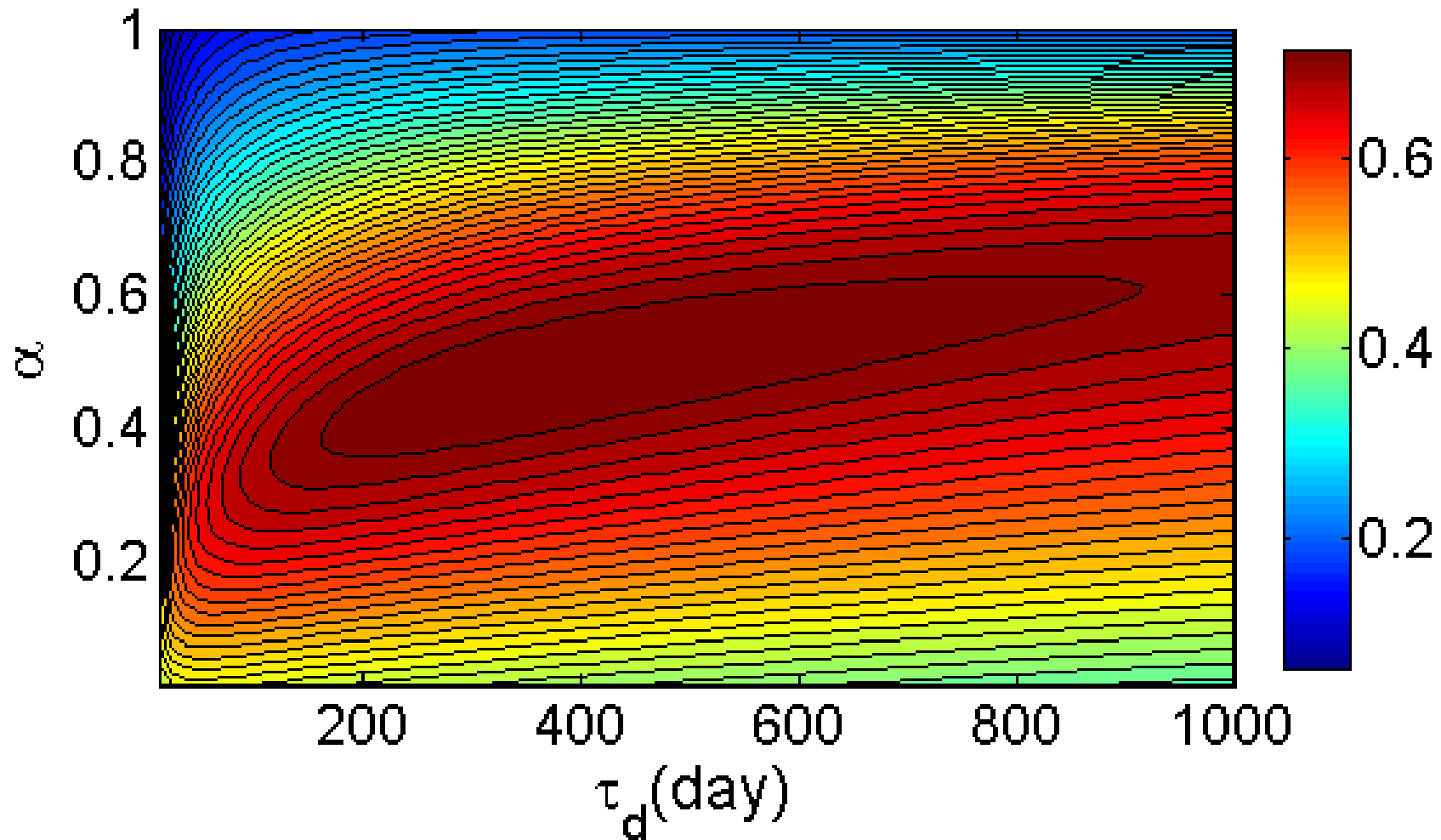
Linear Correlation of T_i with Dk_i



$$\alpha = 0.5, \quad \tau_d = 2.6 \text{ yrs}$$

physrev

Linear Correlation of T_i with Dk_i



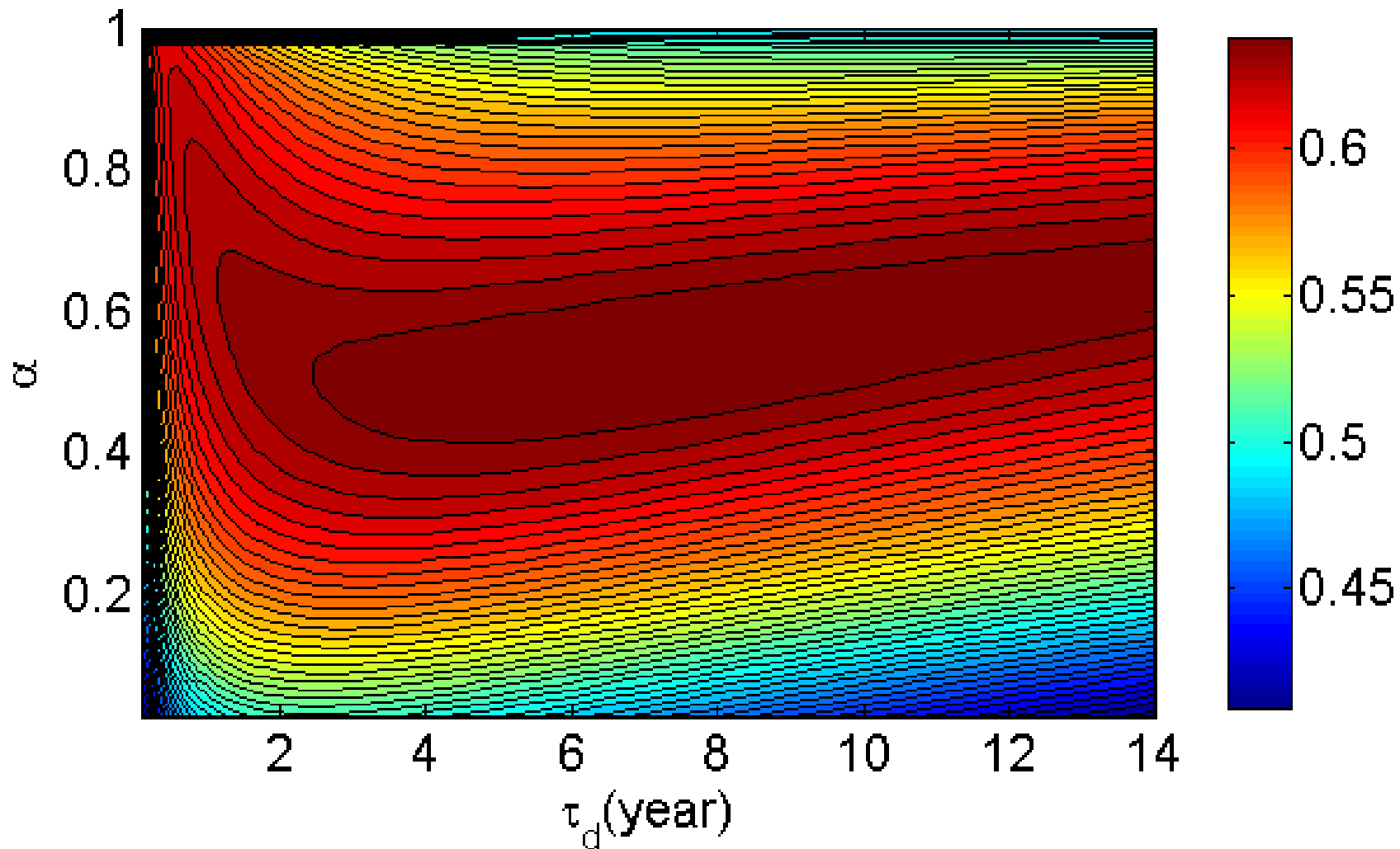
$$\alpha = 0.48, \quad \tau_d = 363 \text{ days}$$

hep-th

What if T_i isn't linear with Dk_i ?

- The previous correlation contour plots rely on the assumption of linearity between real traffic and recent citations.
- Can we relax this assumption to something more reasonable?
- Assume monotonic relationship only
- There is a correlation measure adapted for such a situation: Spearman Rank Correlation
 - Changes in Dk_i that do not lead to rank changes will not affect the correlation.
 - We should expect peaks that are broadened due to this decrease in sensitivity.

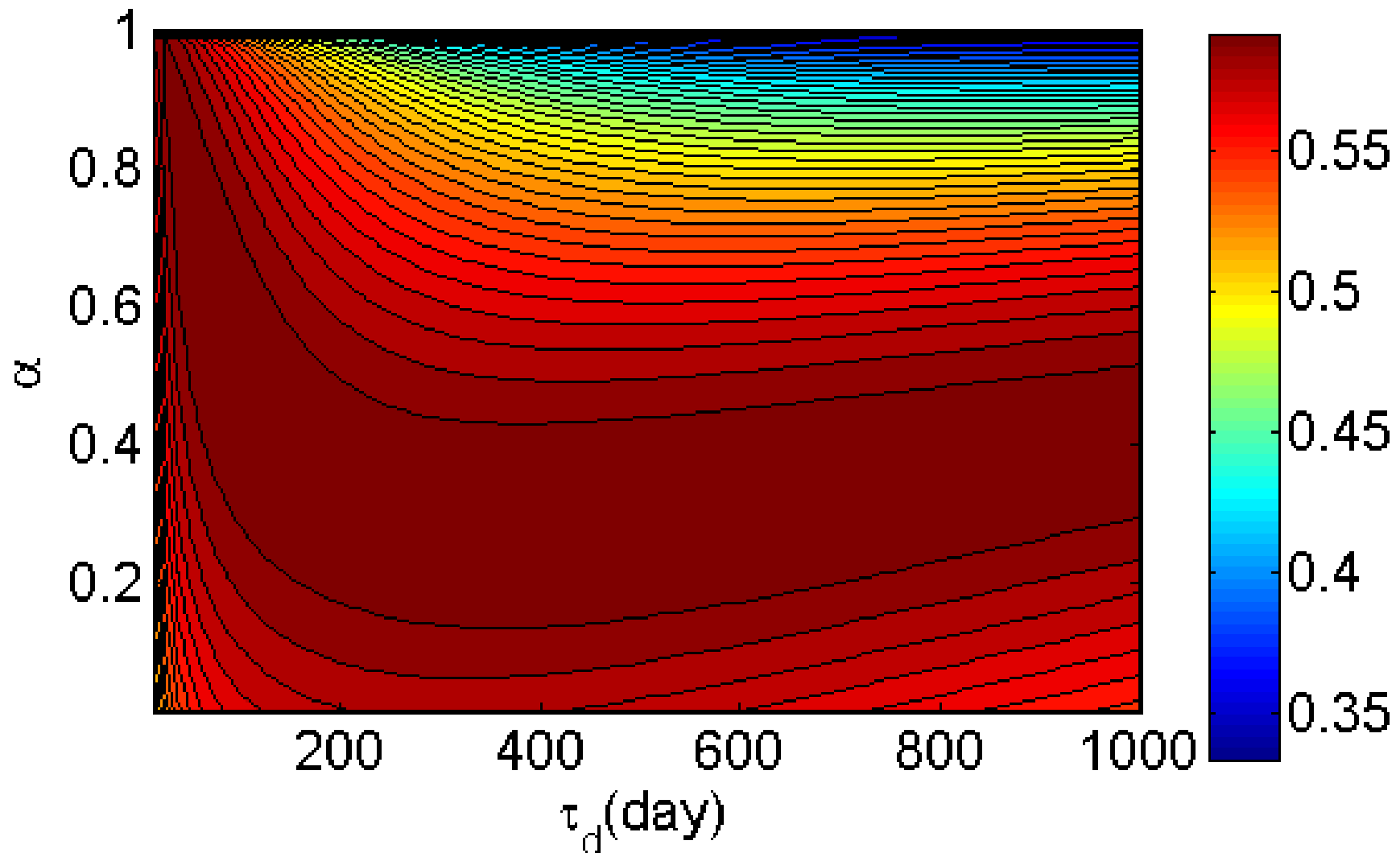
Rank Correlation of T_i with Dk_i



$\alpha = 0.55, \quad \tau_d = 8 \text{ yrs}$

physrev

Rank Correlation of T_i with Dk_i



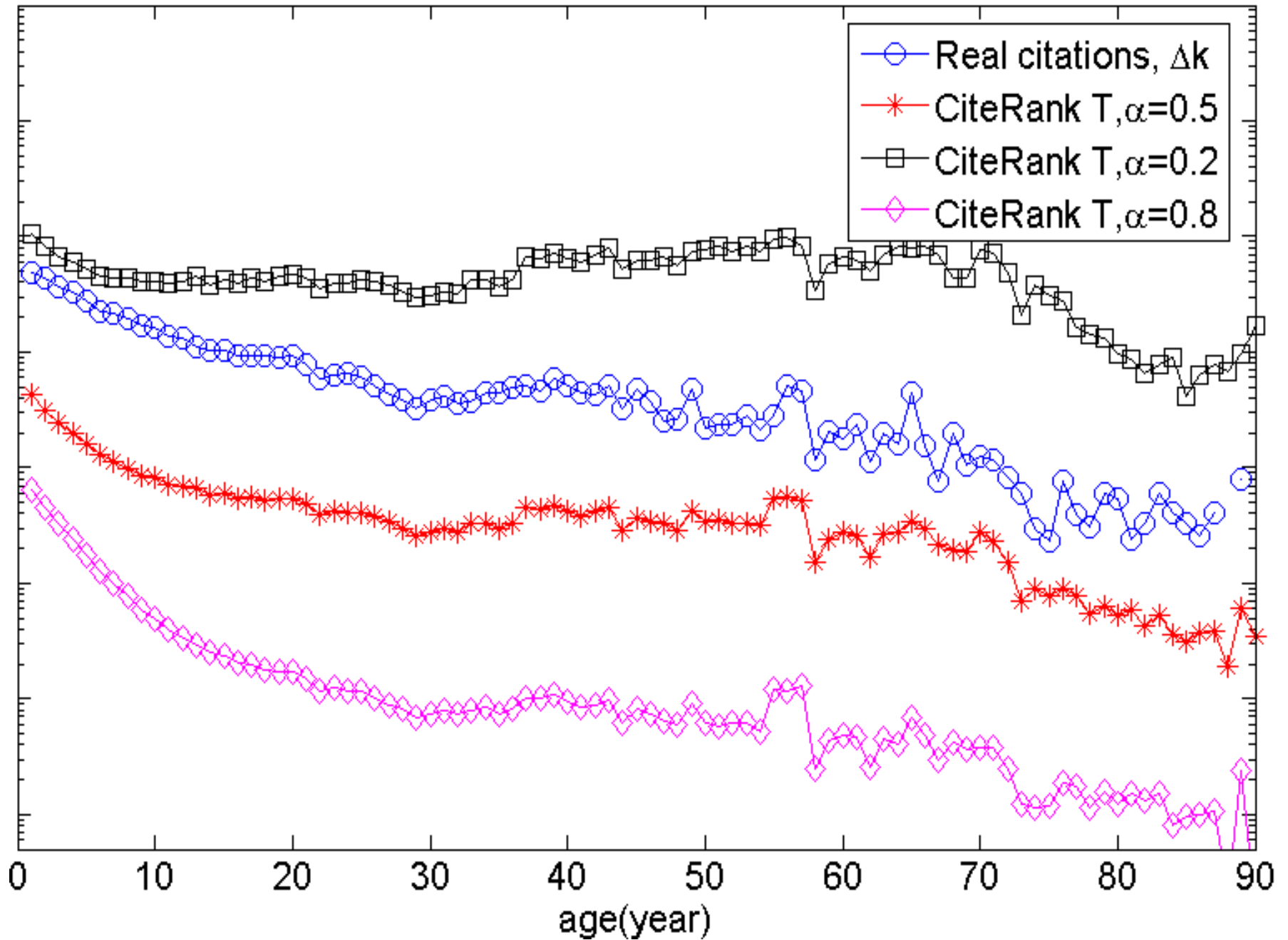
$$\alpha = 0.313, \quad \tau_d = 525 \text{ days}$$

hep-th

Correlation from Age Distribution

- Why is the peak correlation attained at those values of the parameters?
- In what way is traffic prediction getting better?
- Look at linear correlation for physrev
 - Take the slice $t_d = 2.6$ yrs (optimal) and look at effect of varying a .
 - Examine the average age distribution:
 - Real citations , Dk_i
 - Predicted traffic , T_i

Age Distribution



Concluding Remarks

- Good agreement in estimation of \mathbf{a} over networks:

$$\alpha = 0.50, \quad \tau_d = 2.6 \text{ yrs}$$

- On average, the typical researcher follows citation chain of length ~ 2
 $\alpha = 0.48, \quad \tau_d = 363 \text{ days}$

- Future explorations

Precise relation between Dk_i and T_i^{real}

Sampling of actual traffic

Acknowledgements

- Support:

Brookhaven National Lab, Division of Material Science,
U.S. Department of Energy

- Collaborators:

S. Maslov, S. Redner, H. Xie, Y. Koon-Kiu, P. Chen

- Thanks to:

Mark Doyle, Marty Blume, Paul Dlug of the Physical
Review Editorial Office



Citing Age Distribution

$T(t)$ ~ traffic from CiteRank model as a function of age

$T(t)$ is comprised of two varieties of traffic:

Direct traffic ~ $T_d(t)$ – arrive at paper via initial selection

Indirect traffic ~ $T_i(t)$ – arrive at paper via citation

$$T_i(t) = (1 - \alpha) \int_0^t T(t') P_c(t, t') dt'$$

$P_c(t, t')$ ~ fraction papers of age t' → papers age t

To good approximation:

$$P_c(t, t') \sim \frac{e^{-(t-t')/\tau_c}}{\tau_c}$$

or in fourier space:

$$T(\omega) = \frac{T_d(\omega)}{1 - (1 - \alpha)/(1 + i\omega\tau_c)}$$

Then, for the tail of $T(t)$, an exp. fit can be made with:

$\tau_{CR} = \tau_d + \tau_c(1 - \alpha)/\alpha$ so, insisting this tail fit real traffic:

$$(2 - 1/\alpha) = \tau_d / \tau_c$$