Network routing

CS 240

Is article popularity in Wikipedia contagious?

Presented by:

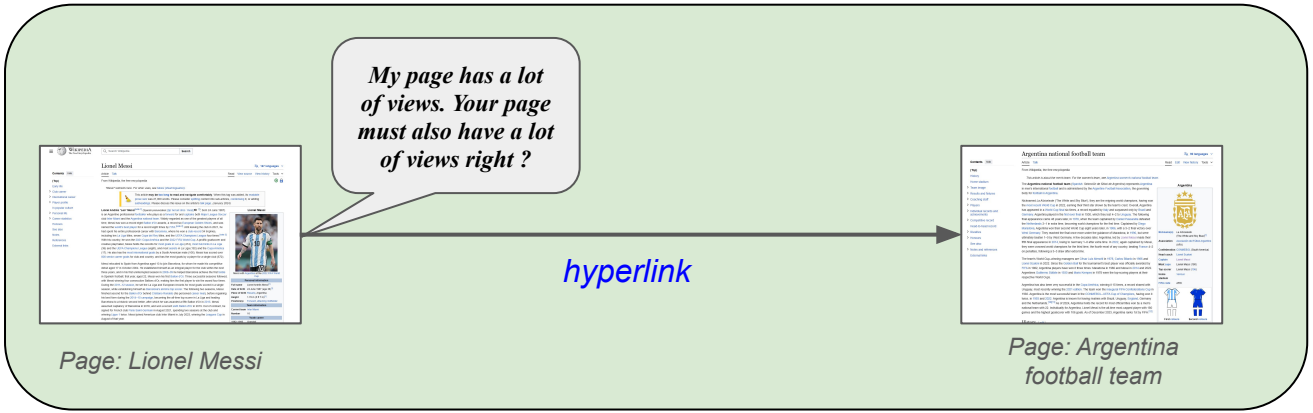
Team Hermes

Nishanth Prasanna Kumar, Noah Tsai, Prathik Somanath, Shubhank Joshi, Vivek Venkateshprasad

The Problem

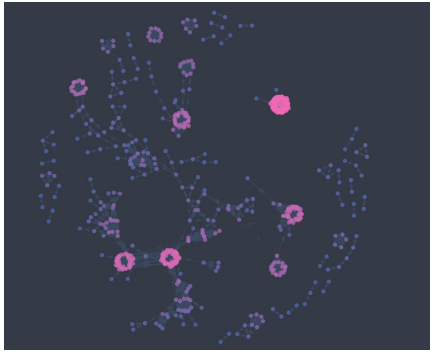
Are Wikipedia articles hyperlinked by popular* pages, more popular ?

**popularity in terms of page views*



Basically, are you cool if your neighbours are cool ?

The Solution



One of the graphs we generated using the “SAHS” dataset

1

Make a wikipedia digraph with the nodes as the articles and their hyperlinks as the edges

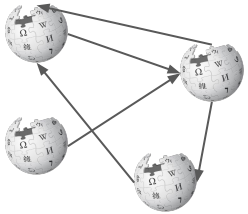
2

Measure the “importance” of these nodes and assign it a score

3

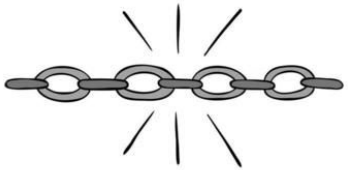
Compare this score with the actual views and make inferences

Motivation



Analysis on the structure of Wikipedia

How much of an impact does a hyperlink have



Novelty and curiosity

~~To get a good grade in network routing~~

Previous work

Many popular works done on wikigraphs focus on click stream datasets rather than take into consideration the graph as a whole.

https://dl.acm.org/doi/abs/10.1145/3038912.3052613
https://link.springer.com/chapter/10.1007/978-3-319-47602-5_41

Takes into consideration

- Location of the link.
- Content of the page.
- Context of the link.

Great papers to study about human behaviour.

Not so great to study the properties of influence due to the nature of a graph.

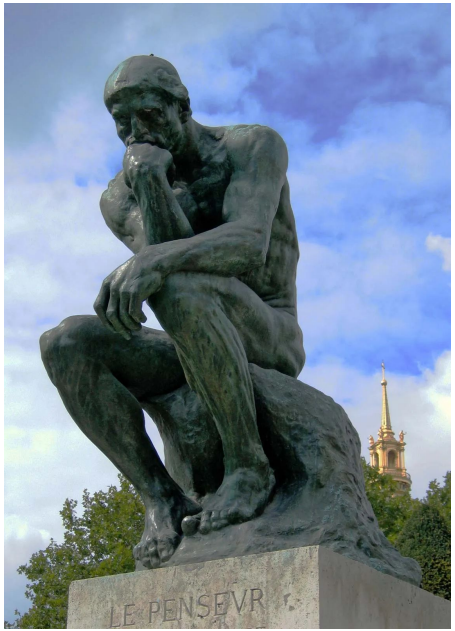
Previous work

Majority of the people access wikipedia through a search engine (Almost 66%). But, there is a significant amount of views through a click network.

https://firstmonday.org/ojs/index.php/fm/article/view/1765/1645
https://en.wikipedia.org/wiki/Wikipedia:Google_statistics#:~:text=Wikipedia%20derives%2066%25%20of%20traffic,a%20quarter%20of%20total%20traffic).

So … We need to pick our dataset carefully ! Ones that are not accessed by search engines as frequently

Previous work



All roads lead to Rome,
or Philosophy in this case !
These papers give important insights on popularity in terms of “Edits”

https://pdodds.w3.uvm.edu/research/papers/ibrahim2017a/
https://ieeexplore.ieee.org/abstract/document/8307100
e0190674. https://doi.org/10.1371/journal.pone.0190674

But does Rome lead to all roads ?

How important are the roads between cities ?

How do we even measure the importance of these nodes ?
These papers explain how

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7351682/
https://link.springer.com/chapter/10.1007/978-3-319-47602-5_41
https://dl.acm.org/doi/abs/10.1145/3038912.3052613

Contribution - Data scraping

We scraped our own data of different categories.

Namely: “South American history stubs”, “Indian Artists”,”Crocodilian” categories

Why ?

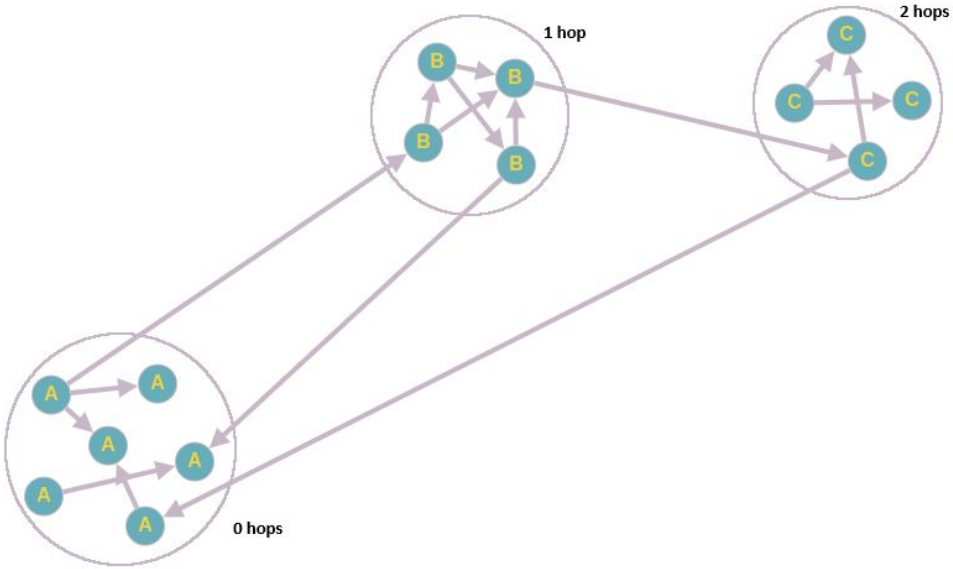
Available datasets do not structure the data in terms of the category.

We needed to structure our data in terms of the category it belongs to and the hops to the neighbours outside it’s category.

How ?

We used the wikimedia API and the wikipedia page views API

It took over 2 weeks !!!!



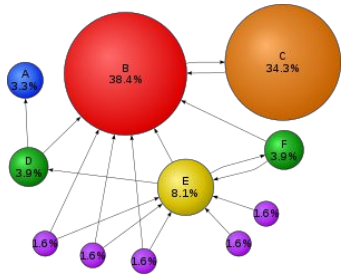
Contribution- Calculating importance

How do we find out the importance of nodes ?

Betweenness centrality: which node acts most like a “bridge”

Katz centrality: which node has most immediate neighbours

PageRank : counting the number and quality of links to a page



https://en.wikipedia.org/wiki/PageRank

But !!!

*None of these importance measures incorporate **page-views** as a metric.*
This means a node with zero views will influence it’s connecting nodes as much as a node with a gazillion views

How do we solve this ?

Contributions - Weighted PageRank

*Introducing **Weighted PageRank***

https://ieeexplore.ieee.org/document/1344743

Damping factor d
How often will somebody randomly stumble on our page

d_{wiki} = 0.9 …. Why ? ↓↓↓

https://dl.acm.org/doi/abs/10.1145/3038912.3052613

$$PR(u) = (1-d) + d \sum_{v \in B(u)} PR(v) W^{in}_{(v,u)} W^{out}_{(v,u)}$$

Edge Weight importance scores W
This part tells us that a more important page(in terms of views) which has lesser edges to/from it - Will be more important

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p \in R(v)} I_p}$$

$$W^{out}_{(v,u)} = \frac{O_u}{\sum_{p \in R(v)} O_p}$$

Contributions

How do we find out if our data is correlated ?

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/correlation-pearson-kendall-spearman/

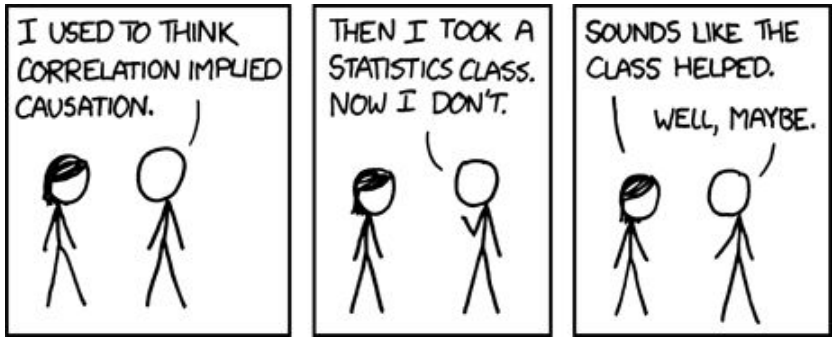
Pearson Correlation: measures linearly related data *Very sensitive to outliers !*

Kendall Correlation: for ordinal data *The data that we have isn't very ordinal*

Spearman Correlation: for less ordinal data
Can measure monotonically related data

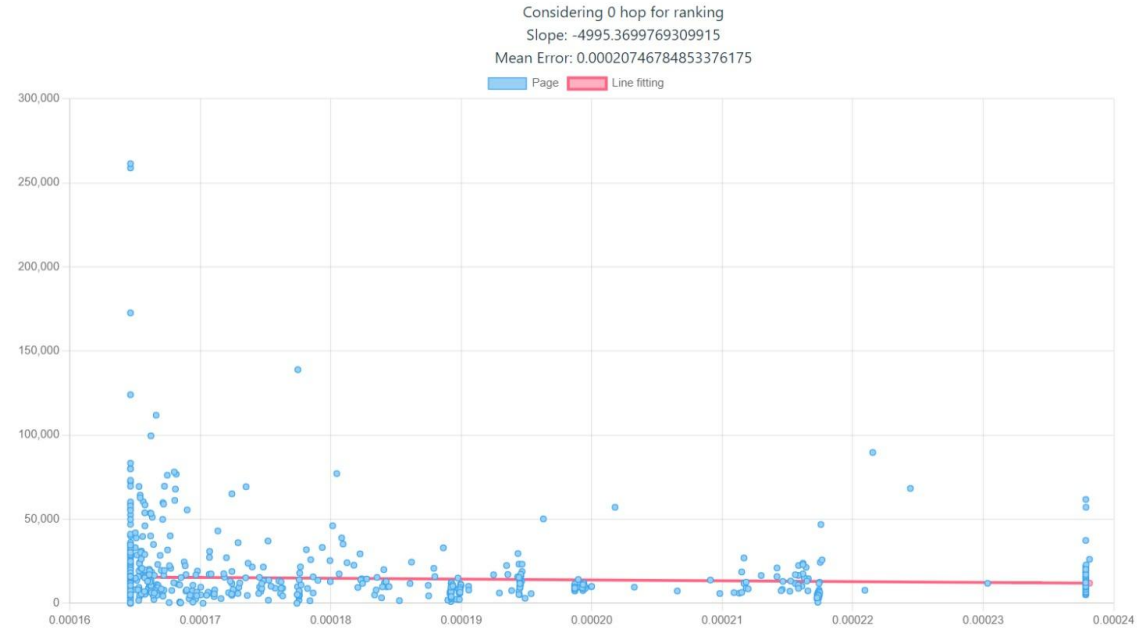
Not sensitive to outliers and our data is cardinal !

For our data,
Spearman looks best !



Experimental Results

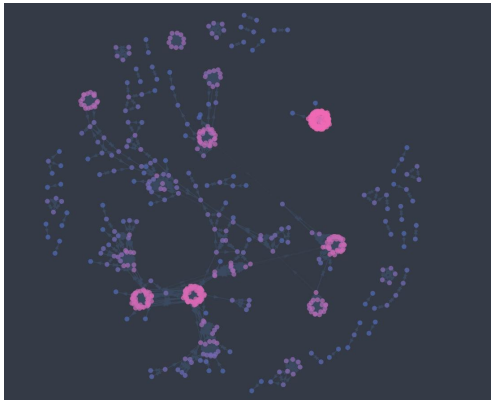
Spearman ρ	Correlation
≥0.70	Very strong relationship
0.40-0.69	Strong relationship
0.30-0.39	Moderate relationship
0.20-0.29	Weak relationship
0.01-0.19	No or negligible relationship



South American History Stubs

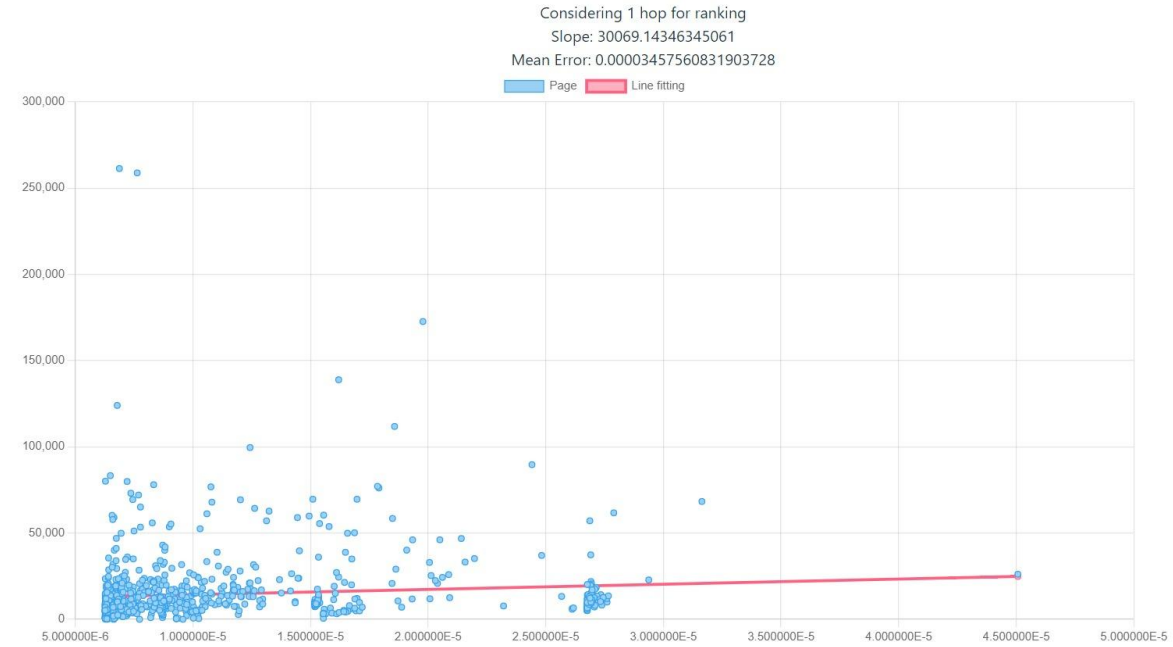
0 hops
(Within Category)

Spearman
rank = **0.16**



Experimental Results

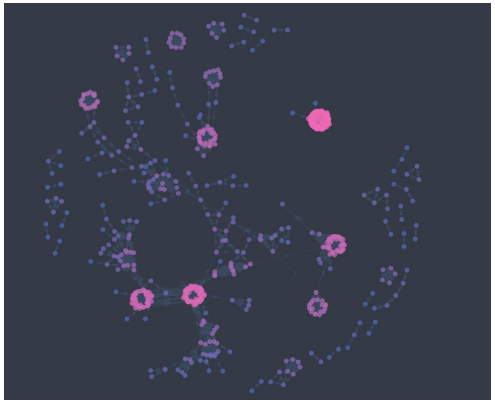
Spearman ρ	Correlation
≥0.70	Very strong relationship
0.40-0.69	Strong relationship
0.30-0.39	Moderate relationship
0.20-0.29	Weak relationship
0.01-0.19	No or negligible relationship



South American History Stubs

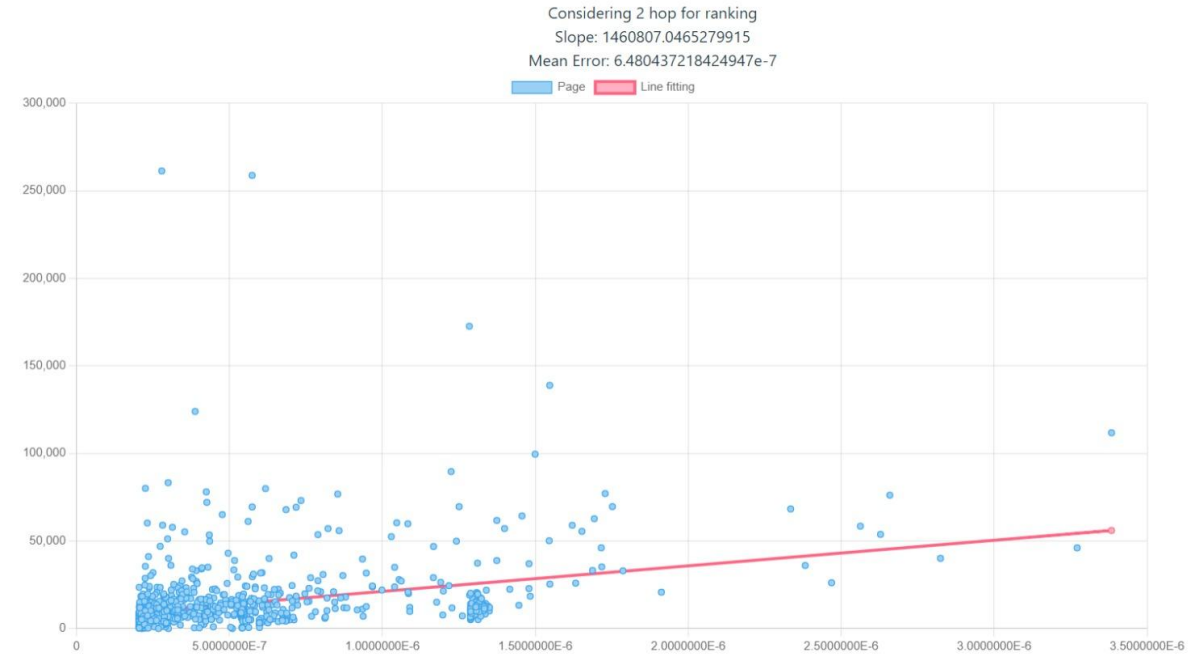
1 hops
(1 hop away from category)

Spearman
rank = **0.38**



Experimental Results

Spearman ρ	Correlation
≥0.70	Very strong relationship
0.40-0.69	Strong relationship
0.30-0.39	Moderate relationship
0.20-0.29	Weak relationship
0.01-0.19	No or negligible relationship

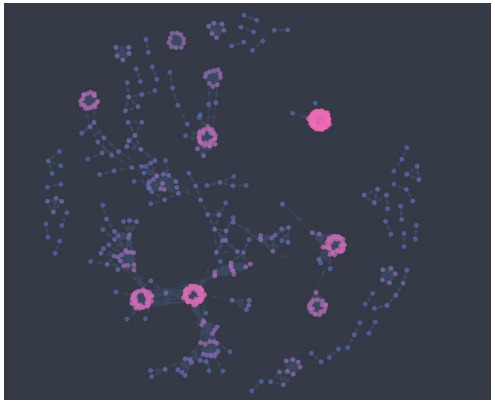


South American History Stubs

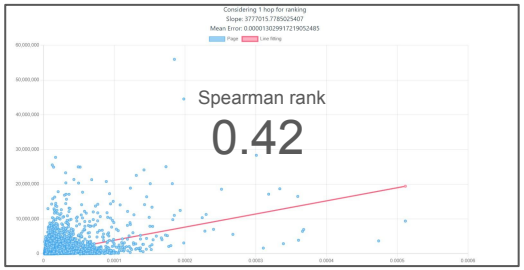
2 hops

(2 hop away from category)

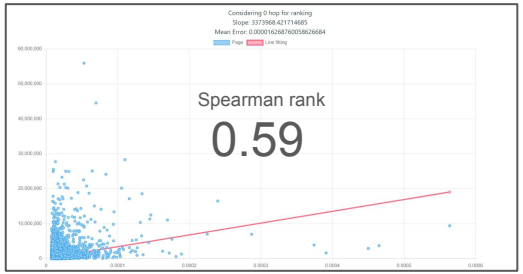
Spearman rank = **0.49**



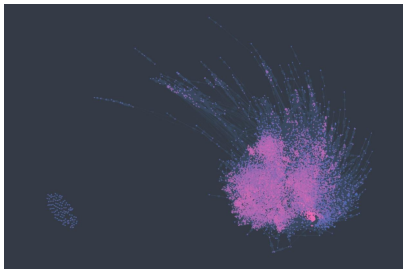
Experimental Results



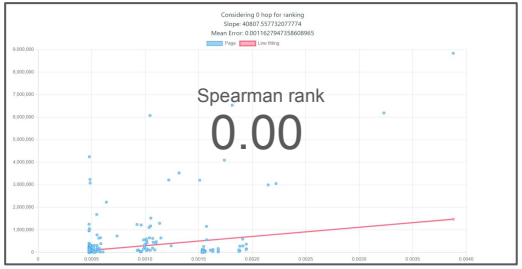
0 hops



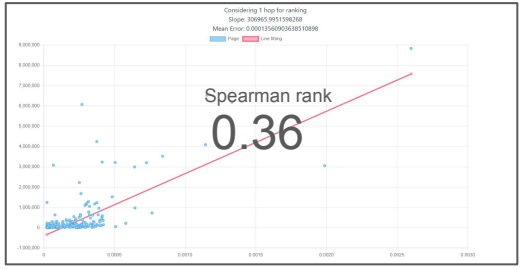
1 hop



Indian Artists



0 hops



1 hop

Crocodilian

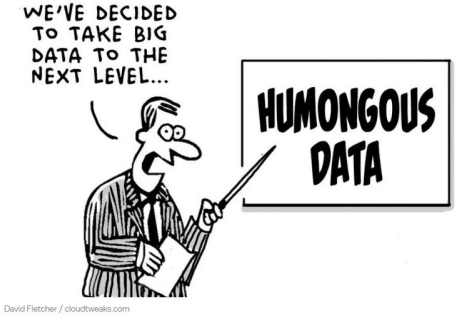
Conclusion

Popularity is indeed contagious in Wikipedia.

We see a great amount of increase in our spearman rank correlation as our hops increase.

We've also noticed categories that are affected less to external searches(Crocodilian, South American History Stubs) tend to have less contagion within a category and we see a rise in popularity contagion as our hops increase.

Future Work



Collect more data to train a machine learning model to predict the views based neighbouring nodes.

Test our current implementation using the hyperlink order to understand the importance of position of the link.

Take into consideration other parameters like the quality of the article(no. of words, well defined structure,no. of edits, etc) while calculating the popularity and rankings.

Thank you

Presented by:

Team Hermes

Nishanth Prasanna Kumar, Noah Tsai, Prathik Somanath, Shubhank Joshi, Vivek Venkateshprasad