EPOC draft:

Page 1:

Good morining ladies and gentlemen. It's Chaoqi here, presenting my work: epoc, a survivial perspective early pattern detection model for outbreak cascades.

I collaborate with Qitian in this work, and our advisor is Doc. Xiaofeng Gao, and Guihai Chen.

Page 2:

In this presentation, we will introduce our problem and the design of our model, finally, we will show part of the experiment's result.

Page3:

Let start by looking at the problems.

Page4:

Every day millions of people express ideas and interact with friends through online platforms like Twitter and facebook. So social medias play a dominant part in people's nowadays life.

What if someone post a piece of infomation, very hot one? How many people will be influenced, what will happen next?

Just see the picture below, on Aug 7th, Musk, CEO of tesla post one message on twitter: he was considering to take Tesla private!

Page5:

See what happened, lots of people went to stock exchange center and buy stocks, forcing the trading system to be shutted down.

And plenty of news and messages emerges from the apps in my cellphone. I guess it is the same in your cellphone, right?

Just one tweet, huge social impact.

Page6:

So if we can predict the popularity of this tweet ahead of time, we can prevent the impart, right?

So problem comes, how can we predict that?

Previous works can be divided into two categories: feature driven and time series.

The feature driven ones, they first extract the feature from the original data, then using classifiers or regressors to get the result.

Time series ones, they use the generative model, see this problem as a counting process or poisson process, then using Max likelihood estimation to handle this.

Page7:

Different from the previous work, we adopt survival theory in this problem, and using cox's extended model to capture the diffusion of tweets.

Page8:

The notion in survival thoery can be simplified as: the longer one creature lives, the low probability it will still be alive, with bunches of factors tuning its survival rate.

Now we adopt this theory into our problems, our assumption is that the longer one tweet exists, the low survival probability (the higher popularity) it will have, with bunches of parameters tuning that.

Page9:

This page shows the real data from Twitter, in figure (a), there are 7 tweet diffusion cascade, the x-axis is time, and y-axis denotes the total retweet number, also called cascade size. Larger size cascade means the popular one.

In figure(b), we draw the retweet count of the the blue one in figure (a) with respect to time t. from t0, it get burst and experience a large retweeting period, right?

Now, if we capture this cascade in survival theory, its life table, the survival rate with respect to time will look like figure(c), during its burst period, the survival rate will drop dramatically, from totally non-popular to totally popular.

Page10:

Then how can we transfer the tweet diffusion to the survival rate decreasing? We use cox's-extended model.

Cox's model look like this, every tweet has a h function, x1,x2,x3 are the time-dependent featues and beta1, beta2, beta3 are parameters, they are shared across each tweeet. Then we use max likelihood estimation to estimate the parameters. To make it simple, I will skip the basic math for cox's model, just present our innovations.

Page11:

OK, we get the survival rates for each tweet with respect to the time now. Pretty nice ha. Here the non-viral means non-popular, viral means popular.

We can see that popular tweets and non-popular tweets will gather, then why not give them a clear boundary? That's what we do next.

For a fixed time t, we plot the survival rate of every tweet like this. (next page)

Page12:

After that, we use two Gaussian distribution to fit the non-viral and viral class.

Page13:

Then use this formula to get the boundary value at time t, nice formula and the solution, now the S star is the boundary value at time t.

Page14:

In order to make the problem more complete and rigorous, We further veify that the survival boundary is itself a survival curve, for time limit, we will skip here.

Page15:

See what we get here. The red dash line is our survival boundary, given by two gaussian distrubutions.

Page16:

Now, if one new tweet comes,

first, we using cox's-extended model to get its survival curve with respect to time t.

Then whenever this curve goes below the survival boundary, we say the tweet will go popular.

Page17:

We do extensive experiment on two famous dataset, twitter and weibo.

Page18:

Twitter is a very well-known social media, and Weibo is a chinese social platform, just like twitter. All the experimental data are crawlered by API tools.

We compare our model with five baselines from different perspectives, LR(linear regression), SVR(support vector regressio), PreWhether(Beyesian network), SEISMIC(point process), SansNet(survival model).

Our experiments are based on 0.5h, 1h, 1.5h, 2h, 2.5h and 3h.

The evaluation metric is very special, k-coverage and self desiged cost

From this two figure, we can see that the deep green one is our model, and it beat the baselines by a remarkable margin.

Page19:

Another experiment is how much time ahead we can condifently say that this tweet will be popular at the near future, in this experiment, our model also achieve state-of-the-art performance.

Page20:

Thnaks for your listening.