

HOW TO DESIGN AN HONEST RATING SYSTEM

Sergey I. Nikolenko^{1,2}

AI Rush 2017

Dnipro, February 18, 2017

¹Laboratory for Internet Studies, NRU Higher School of Economics, St. Petersburg

²Steklov Institute of Mathematics at St. Petersburg

Random facts:

February 18, 1268: forces of the Livonian Order defeated by Dovmont of Pskov in the Battle of Rakvere

February 18, 1930: Ellie Farm Ollie became the first cow to fly and be milked inside an aircraft

February 18, 1943: Joseph Gebbels delivers his *Sportpalast* speech

February 18, 1954: the first Church of Scientology was established in Los Angeles

BAYESIAN RATING SYSTEMS

MY PERSONAL MOTIVATION

- «What? Where? When?»: a team game of answering questions. Sometimes it looks like this...



MY PERSONAL MOTIVATION

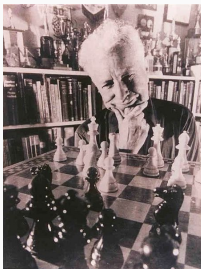
- ...but usually it looks like this:



- Teams of ≤ 6 players answer questions.
- Whoever gets the most correct answers wins.
- My motivation was to create a rating system that would predict tournament results by team rosters.
- Characteristic features that make the problem hard:
 - it's a hobby: players have no contracts, teams do not have permanent rosters, playing for many teams is common;
 - hence, we cannot just make a rating list of the teams, we need to go deeper, to individual players;
 - but we do not know how players do, only team results;
 - relatively few questions per tournament (36, 45, 60), hence multiway ties;
 - undersized teams are common.

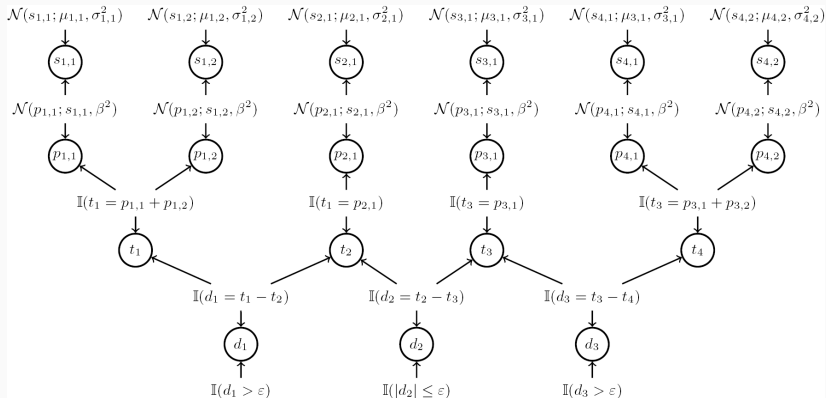
INTRODUCTION

- In probabilistic rating models, Bayesian inference aims to find a linear ordering on a certain set given noisy comparisons of relatively small subsets of this set.
- Useful whenever there is no way to compare a large number of entities directly, but only partial (noisy) comparisons are available.
- We will stick to the metaphor of matches and players.
- Elo rating system: first probabilistic rating model.



- Bradley–Terry models: assume that each player has a “true” rating γ_i , and the win probability is proportional to this rating: γ_1 wins over γ_2 with probability $\frac{\gamma_1}{\gamma_1 + \gamma_2}$.
- Inference: fit this model to the data from matches played.
- Several extensions, but large matches are hard for Bradley–Terry models.
- The model that looked right to us for «What? Where? When» was TrueSkill.

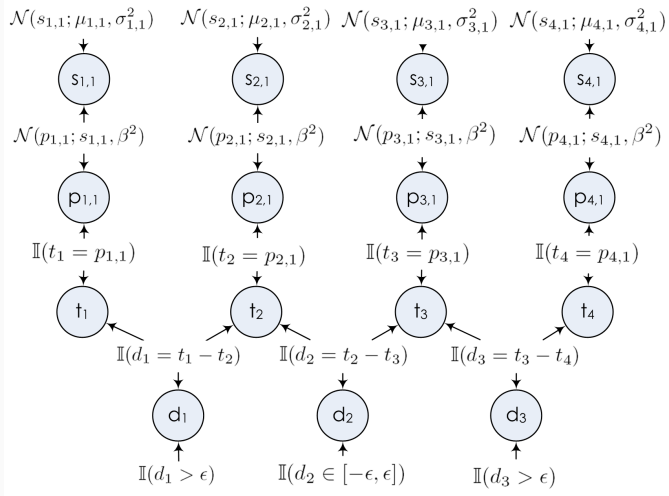
TRUESKILL FACTOR GRAPH



- TrueSkill was initially developed in MS Research for Xbox Live gaming servers [Graepel, Minka, Herbrich, 2007].
- Given results of team competitions, learn the ratings of players of these teams.
- Direct application – matchmaking: find interesting opponents for a player or team.
- [Graepel et al., 2010]: AdPredictor. Predicts CTRs of advertisements based on a set of features: the features are a team, and the team wins whenever a user clicks the ad.
- Basic idea: construct a probabilistic graphical model for a tournament.

- There is no evidence per se, it is incorporated in the structure of the graph, we just have to marginalize by message passing.
- The marginalization problem is complicated by the step functions at the bottom; solved with Expectation Propagation [Minka, 2001]:
 - approximate messages from $\mathbb{I}(d_i > \epsilon)$ and $\mathbb{I}(|d_i| \leq \epsilon)$ to d_i with normal distributions;
 - repeat message passing on the bottom layer of the graph until convergence.

EXAMPLE: A MATCH OF FOUR PLAYERS



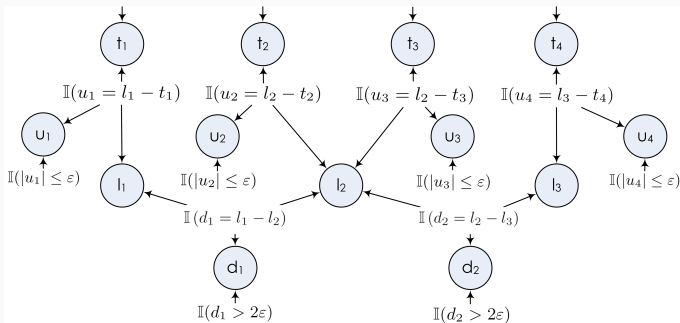
- TrueSkill looked perfect for «What? Where? When».
- But it didn't really work due to the following properties of the «What? Where? When» dataset.
 1. Teams vary in size (max 6 players, but often incomplete):
 - undersized teams stand a very good chance against a full one,
 - so adding player performances to get the team performance does not work.
 2. Large multiway ties are common; 30–40 different places (35–50 questions) in a tournament with a thousand teams:
 - this is deadly for TrueSkill: consider four teams with performances p_1, \dots, p_4 , 1 has won, and 2–4 drew behind it;
 - then the factor graph tells us that

$$t_2 < t_1 - \epsilon, \quad |t_2 - t_3| \leq \epsilon, \quad |t_3 - t_4| \leq \epsilon.$$

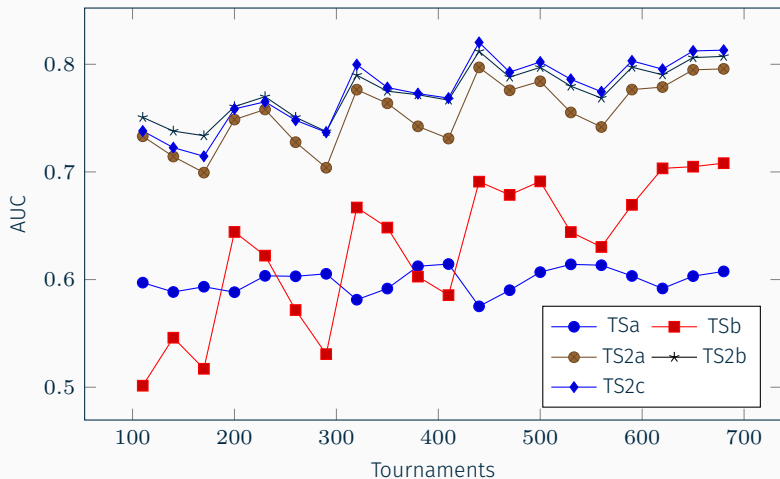
- t_3 may actually nearly equal t_1 , and t_4 may exceed t_1 !

CHANGES IN THE FACTOR GRAPH

- For the multiway tie problem, we add another layer in the factor graph, namely the layer of *place performances* l_i .
- Each team performs in the ϵ -neighborhood of its place performance, and place performances relate to each other with strict inequalities like $l_2 < l_1 - 2\epsilon$.
- Then it's inference as usual, no slowdown in convergence.



EXPERIMENTAL RESULTS



Average AUC over a sliding window of 50 tournaments.

**MORE DETAILED DATA
LEADS TO A SIMPLER MODEL**

- Several years ago, «What? Where? When?» tournament database started collecting question-wise data.
- That is, we now know which specific questions a team has answered; previously we only had standings in a tournament.
- So when I got back to the problem of «What? Where? When?» ratings, I found the problem greatly simplified.

- Sample relevant application:
 - consider a test suite with many questions that tests something (e.g., IQ or a specific);
 - participants answer a random subset of questions;
 - we need to rate participants but questions are different, so the complexity level cannot be perfectly balanced.
- «What? Where? When?» is just like that, but participants are working on the test in teams.

- Baseline model – logistic regression; we model:
 - each player i with his or her skill s_i ,
 - each question q with its complexity score c_q ,
 - add the global average μ ,
 - and train the logistic model

$$p(x_{tq} \mid s_i, c_q) \sim \sigma(\mu + s_i + c_q)$$

for each player $i \in t$ of a participating team $t \in \mathcal{T}^{(d)}$ and each question $q \in Q^{(d)}$, where $\sigma(x) = 1/(1 + e^x)$ is the logistic sigmoid, and x_{tq} denotes whether team t answered question q correctly.

- The logistic model basically assumes that each player successfully answered every question that the team had answered.
- But in fact we do not know which player or players have answered.
- We only can assume that if the team has failed then no one from this team has done it.
- This situation is similar in spirit to presence-only data models found in, e.g., ecology [Ward et al., 2009; Royle et al., 2012].

- Hence, a model with latent variables.
- For each player-question pair, we add a latent variable z_{iq} which means «player i has answered question q ».
- For these variables, we have the following constraints:
 - if $x_{tq} = 0$ then $z_{iq} = 0$ for every player $i \in t$;
 - if $x_{tq} = 1$ then $z_{iq} = 1$ for at least one player $i \in t$.

- Model parameters are still skill and complexity of the tasks:

$$p(z_{iq} \mid s_i, c_q) \sim \sigma(\mu + s_i + c_q).$$

- Training with EM:
 - E-step: fix all s_i and c_q , compute expected values of latent variables z_{iq} as

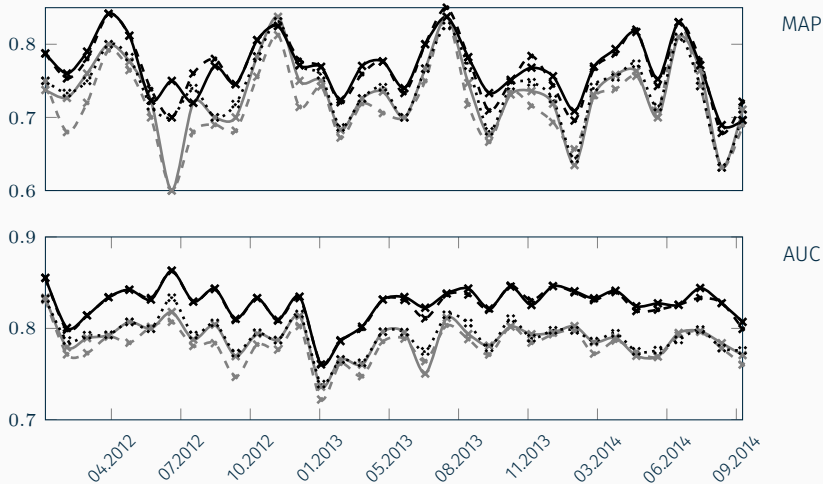
$$\mathbb{E}[z_{iq}] = \begin{cases} 0 & \text{if } x_{tq} = 0, \\ p(z_{iq} = 1 \mid \exists j \in t \ z_{jq} = 1) = \frac{\sigma(\mu + s_i + c_q)}{1 - \prod_{j \in t} (1 - \sigma(\mu + s_j + c_q))}, & \text{if } x_{tq} = 1; \end{cases}$$

- M-step: fix $\mathbb{E}[z_{iq}]$ and train the logistic model

$$\mathbb{E}[z_{iq}] \sim \sigma(\mu + s_i + c_q).$$

RESULTS

- And, sure enough, it works fine.



Рейтинг ЧГК

игроки

команды

турниры

вопросы

Рейтинг-лист игроков

РЕЛИЗ:

ЯНВАРЬ 2015

РЕЙТИНГ-ЛИСТ ИГРОКОВ НА ЯНВАРЬ 2015

Показывать по 100 записей

Введите id или начало фамилии:

Место	id	Фамилия	Имя	Отчество	Команда	Сыграно	Взято	Рейтинг
1	27177	Ромашова	Вероника	Михайловна	ЛКИ	5278	3784	545.753
2	3083	Белявский	Дмитрий	Михайлович	ЛКИ	4831	3396	543.520
3	27403	Руссо	Максим	Михайлович	ЛКИ	7180	5205	534.540
4	4270	Брутер	Александра	Владимировна	ЛКИ	8244	5974	533.405
5	18332	Либер	Александр	Витальевич	Рабочее название	8658	6210	532.921
6	1585	Архангельская	Юлия	Сергеевна	Ксеп	7542	5286	531.866
7	24384	Пашковский	Евгений	Александрович	ЛКИ	7552	5374	531.327
8	8333	Губанов	Антон	Александрович	Команда Губанова	4475	3153	530.871
9	16332	Крапиль	Николай	Валерьевич	Ксеп	6927	4899	530.559
10	21487	Моносов	Борис	Яковлевич	Команда Губанова	5115	3645	530.175

Рейтинг ЧГК

ИГРОКИ

КОМАНДЫ

ТУРНИРЫ


ВОПРОСЫ

Александр Друзь vs. Максим Поташев

Ссылка на сайт рейтинга МАК

Друзь


Александр Абрамович



Ссылка на сайт рейтинга МАК

Поташев

Максим Оскарович



ИСТОРИЯ РЕЙТИНГА

ИСТОРИЯ ТУРНИРОВ

ИСТОРИЯ РЕЙТИНГОВ

Рейтинг	Команда	Место	Дата	Место	Команда	Рейтинг
450.138	Трансфера	303	Январь 2015	31	Афина	514.643
446.669	Трансфера	336	Декабрь 2014	29	Афина	514.697
442.559	Трансфера	348	Ноябрь 2014	25	Афина	513.155
437.999	Трансфера	400	Октябрь 2014	25	Афина	512.083
445.791	Трансфера	306	Сентябрь 2014	23	Афина	515.737
441.473	Трансфера	372	Август 2014	23	Афина	516.513
446.730	Трансфера	284	Июль 2014	22	Афина	516.755
448.598	Трансфера	322	Июнь 2014	24	Афина	516.603

Thank you for your attention!

Final takeaway points:

- Try to collect new data!

The new model is much simpler than TrueSkill but still works better because we have more detailed data available.

- Don't be afraid to work on your passions!

If you are excited about the problem, you will make better progress, and «real» applications will find you.