A New Bayesian Rating System for Team Competitions

Sergey Nikolenko[1,2] and Alexander Sirotkin[1,3]

[1]St. Petersburg Academic University – Nanotechnology Research and Education Centre of the RAS, 194021, St. Petersburg, Khlopina 8, korp. 3
[2]Steklov Mathematical Institute, 191023, St. Petersburg, Russia, nab. r. Fontanka, 27
[3]St. Petersburg Institute for Informatics and Automation of the RAS, 199178, St. Petersburg, 14 Line V.O., 39

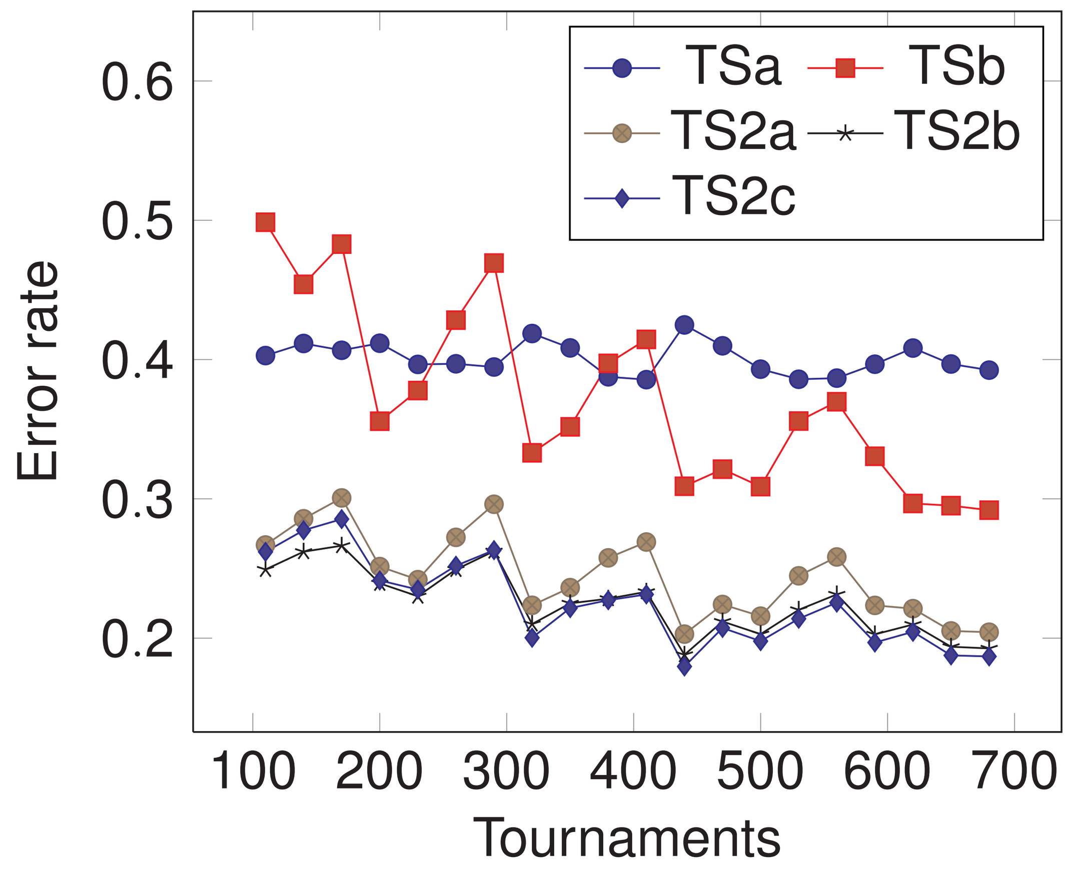
Introduction

- 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.
- Elo rating, Bradley–Terry models, and recently TrueSkill™ [Graepel, Minka, Herbrich, 2007].
- TrueSkill™ was initially developed in Microsoft Research for Xbox 360 gaming servers.
- Applications: matchmaking, AdPredictor, etc.

Model variables

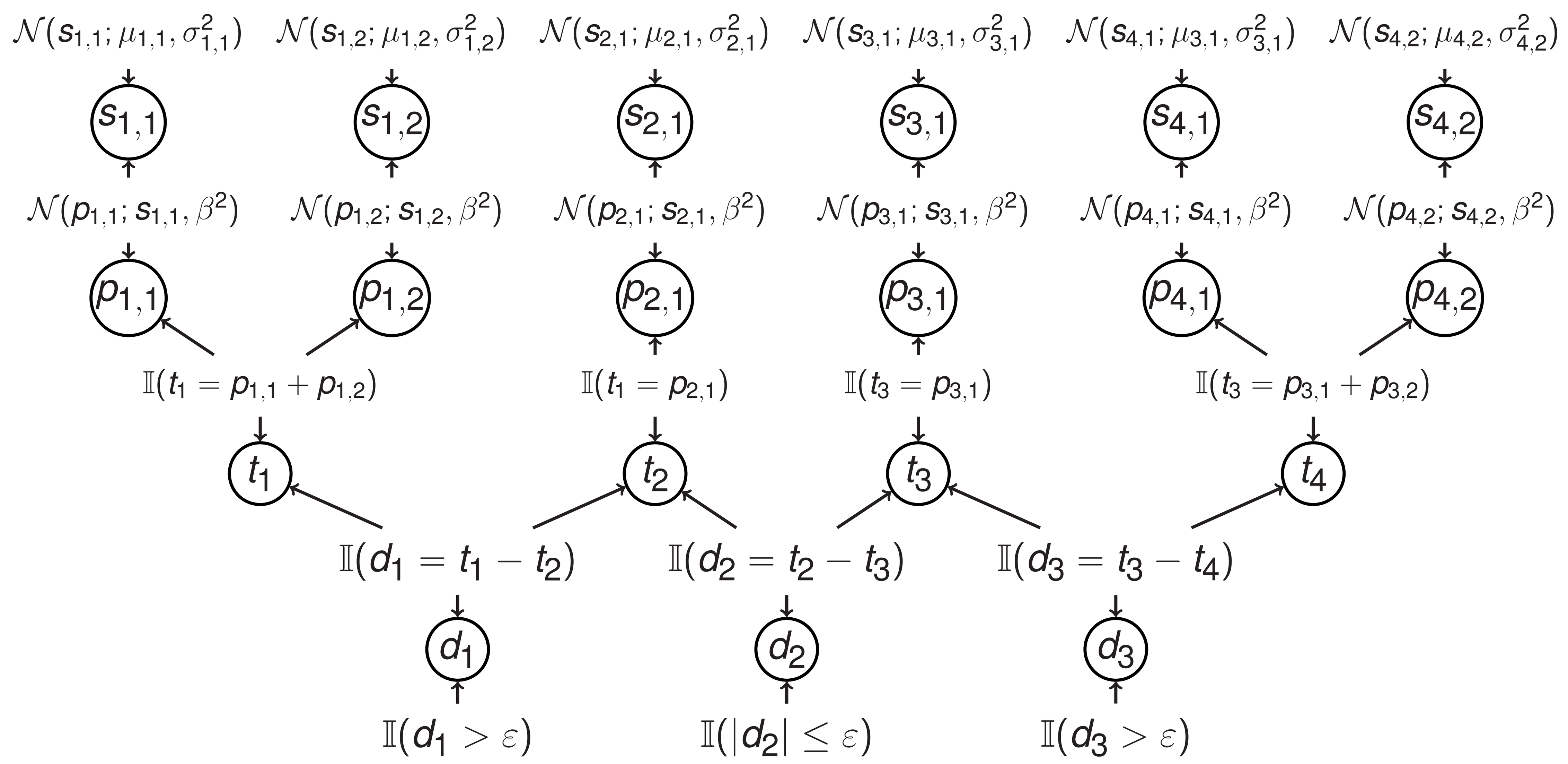
- Layers of TrueSkill factor graph:
  - $s_{i,j}$ – skill of player i from team j; normally distributed around $\mu_{i,j}$ with variance $\sigma_{i,j}$;
  - $p_{i,j}$ – performance of player i from team j;
  - t_j – performance of team j;
  - d_j – difference in performance between teams who took neighboring places in the tournament; a tie corresponds to $|d_j| \le \varepsilon$; a win, to $d_j > \varepsilon$;
  - our contribution: l_j – place performance; u_j – difference between team performance and the corresponding place performance, $|u_j| \le \varepsilon$.
- Inference is complicated by indicator functions at the bottom; solved with Expectation Propagation [Minka, 2001].

Experimental Results



Average error rate over the sliding window of 50 tournaments.

TrueSkill factor graph

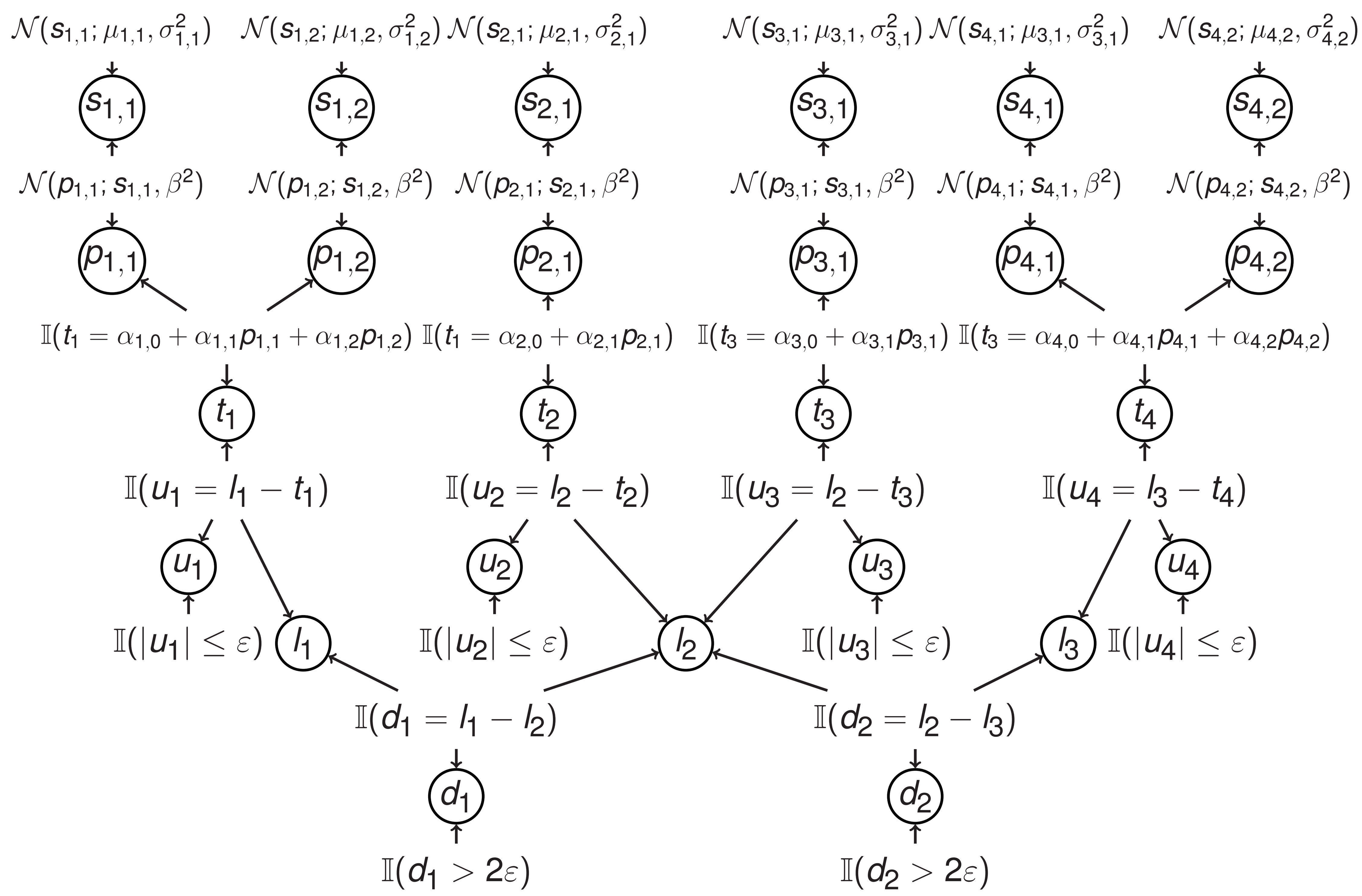


A sample TrueSkill™ factor graph: four teams, teams 1 and 4 have two players each; teams 2 and 3, one player each. Team 1 won, teams 2 and 3 drew behind it, team 4 placed last.

TrueSkill problems

- Large multiway ties are deadly for TrueSkill™. Consider four teams in a tournament with performances $p_1, \ldots, p_4$.
- Team 1 has won, teams 2–4 drew behind.
- Then the factor graph tells us that
$$p_2 < p_1 - \epsilon, \quad |p_2 - p_3| \le \epsilon, \quad |p_3 - p_4| \le \epsilon.$$
- Team 3's performance may actually nearly equal p_1, and p_4 may exceed p_1!
- Moreover, these boundary cases are realized in practice when unexpected results occur.
- Another undesired feature of TrueSkill™ is the assumption that a team's performance is the sum of player performances: in many competitions, an undersized team stands a very good chance against a full one.

Our factor graph



Our factor graph for the same case.

Team performance functions

- We can easily use any affine function for team performance, e.g., average.
- To approximate nonlinear functions, replace player performances with their estimates provided by the prior ratings μ_i. E.g., to approximate $t = p_1^2 + p_2^2 + \ldots + p_n^2$ we replace it with
$$t = \mu_1 p_1 + \mu_2 p_2 + \ldots + \mu_n p_n$$
(here p_i are model variables, and μ_i are constants fixed before inference).
- For our dataset, the function that worked best was (TS2b and TS2c on the graph)
$$t_i = \begin{cases} \frac{\sum_{j=1}^{n_i} p_{i,j}}{n_i} \cdot (0.88 + 0.02 n_i), & n_i \le 6, \\ \sum_{j=1}^{n_i} p_{i,j} \cdot \frac{\sum_{j=1}^{6} \mu_{i,j}}{6 \sum_{j=1}^{n_i} \mu_{i,j}}, & n_i > 6, \end{cases}$$
where n_i is number of players in team i.
- Obviously, it wouldn't work for other applications, so please tune it yourself.

Mail: snikolenko@gmail.com, sergey@logic.pdmi.ras.ru

WWW: http://logic.pdmi.ras.ru/~sergey