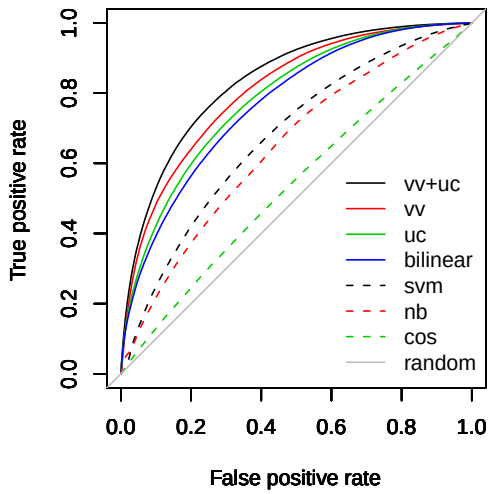

Figure 1: ROC curves of different models

Method	AUC	P@5	P@10	P@20
vv+uc	0.8360	0.9152	0.9079	0.8942
vv	0.8090	0.8810	0.8807	0.8727
uc	0.7857	0.9046	0.8921	0.8694
bilinear	0.7701	0.9028	0.8894	0.8668
svm	0.6768	0.7814	0.7678	0.7497
nb	0.6465	0.7660	0.7486	0.7309
cos	0.5382	0.6834	0.6813	0.6754

Table 2: AUCs and precisions of different models.

rather than in isolation. Next, **uc** outperforms **bilinear** (significantly in AUC, P@10 and P@20), showing per-user and per-comment latent factors help. Note that **vv** outperforms **uc** in ROC, AUC and P@20, but is worse than **uc** in P@5 and P@10; we will take a closer look at this later. Finally, the full model **vv+uc** significantly outperforms both **vv** and **uc**, achieving 0.83 in AUC, and close to 90% in precision at rank 20.

4.3.2 Break-down by user activity level

Next, we investigate model performance in different subsets of the test set. For succinctness, we use AUC as our performance metric. In Figure 2(a), we breakdown model performance by different author activity levels. In Figure 2(b), we breakdown model performance by different voter activity levels. We also generated similar plots with the y-axis replaced by P@5, P@10 and P@20, and observed the same trend except that **vv** starts to outperform **uc** at different user activity thresholds for different metrics.

Comparison	Metrics	p-value
vv+uc $>$ **vv**	All	$< 10^{-7}$
vv+uc $>$ **uc**	All	$< 10^{-20}$
uc $>$ **bilinear**	All except P@5	< 0.006
bilinear $>$ **svm**	All	$< 10^{-20}$
vv $>$ **svm**	All	$< 10^{-20}$
svm $>$ **nb**	All	$< 10^{-8}$
nb $>$ **cos**	All	$< 10^{-20}$

Table 3: Paired t-test results. Note that **uc** is better than **bilinear** in P@5, but not significant. The orders of **uc** and **vv** are not consistent across different metrics.

Not surprisingly, **vv** performs poorly for raters or authors with no ratings observed in the training data. However, once we have a small amount of ratings, it starts to outperform **uc**, even though intuitively, the textual information in the comment should be more informative than the authorship information alone. Using paired t-tests with significance level 0.05, we report when **vv** starts to significantly outperform **uc** in the following table, which is interpreted as follows — **vv** is not significantly worse than **uc** in metric M if the author of a test comment received at least N_{eq} ratings in the training set, and **vv** significantly outperforms **uc** in metric M if the author received at least N_{gt} ratings in the training set.

Metric M	P@5	P@10	P@20	AUC
# Ratings N_{eq}	50	5	5	5
N_{gt}	1000	50	5	5

Recall that our training/test split is by article. Since we have never observed a rater's preference over the test articles before, it is rather surprising that author information alone can yield 0.8 in AUC score, even for very light authors who have received between 3 and 5 votes in total in the training data. This suggests that users' viewpoints are quite consistent: a large portion of the ratings can be adequately explained by the pair of user identities. One interesting observation is that the number of ratings required for **vv** to outperform **uc** in P@5 is quite high. This suggests that to obtain high precision at the top of a recommended list, comment features are important.

Nonetheless, modeling textual information in addition to author information provides additional improvements. Based on paired t-tests with signifi-