Motivating Dynamic Features for Resolution Time Estimation within IT Operations Management

Kayhan Moharreri, Jayashree Ramanathan, and Rajiv Ramnath
Department of Computer Science and Engineering
The Ohio State University
{moharreri.1, ramanathan.2, ramnath.6}@osu.edu

***Abstract*—Cloud-based services today depend on many layers of virtual technology and application services. Incidents and problems that arise in such complex operational environments are logged as a ticket, worked on by experts and finally resolved. To assist these experts, any machine recommendation method must meet the following critical business requirements: 1) the ticket must be resolved, meeting specific time constraints or *Service Level Targets (SLTs)*, and 2) any predictive assistance must be trustworthy. Existing research uses probabilistic models to recommend transfers between experts based on limited features intrinsic to the ticket content, and does not demonstrate how to meet SLTs. To address this lack of research and ensure SLT-compliance for an incoming ticket given its recommended sequence of experts, there needs to be an accurate time-to-resolve (TTR) estimation. This research aims to identify important features for modeling time-to-resolve estimation given the routing recommendation sequences. This work particularly makes the following contributions: 1) constructs a framework for assessing TTR estimations and their SLT-compliance, 2) applies the assessment to a baseline estimation model to identify the need for better TTR modeling, and 3) uses language modeling to study the impact of anomalous content on the estimation error, and 4) introduces a set of dynamic features, and a methodology to rigorously model the TTR estimation.**

***Index Terms*—Collective Expert Networks; Language Modeling; Regression; Resolution Time Estimation; Service Level Targets; Ticket Resolution; Ticket Routing Recommendations**

I. Introduction

Within today's complex cloud operations made up of layers of technology and application services, customer-perceived problems or incidents often arise in the hundreds or thousands daily. These must be resolved by IT Service Management goals that are critical parts of any service level (SL) agreement:

- *Resolution Goal*: the problem must be resolved by restoring service to the business customer's satisfaction; and
- *SLT Goal*: the resolution process must meet time constraints set by Service Level Targets (SLTs) agreed-to with the customer based on the priority of the ticket.

The Resolution Goal is addressed by logging problems as tickets and then transferring them to the knowledgeable experts (selected from among many) with skills to contribute to the problem resolution. In this area of research a typical representation of experts (i.e. nodes) and their transfers (i.e. edges) is as an expert network [8], [5]. A ticket 'transfer sequence' is a path along the network; machine recommendations aim to reduce transfers needed for resolution.

TABLE I: Glossary: Key Terms

Term	Description
Expert	A technical support team with specialized knowledge and particular set of skills, and responsibilities
Ticket	t a ticket instance with content/attributes, T a ticket set
SLT	Target resolution time defined for each $t \in T$ chosen according to a predetermined priority level.
TRS	Ticket Resolving Sequence of experts for t. R-TRS is a TRS that is labeled as 'Routine'. RecTRS is a recommended TRS (i.e routing recommendation)
TTR	Time To Resolve (i.e. resolution time) of a ticket. Only defined for the resolved tickets.
ETTR	Expected Time To Resolve of a ticket. Used for arriving tickets to estimate their TTR.
MTTR	Mean Time To Resolve computed for T.
MSTR	Mean Steps (i.e. transfers) To Resolve computed for T.

Existing research addresses the Resolution Goal, but the SLT Goal remains unaddressed. The significance comes from the need to improve response times in discovery-oriented environments and make accurate SLT-preserving recommendations. This has also become important for a large emerging class of time-critical applications related not only to Operations Management (i.e. IT service desks); but also to disaster recovery, emergency management, and triage centers.

A. Related Work

Here we first summarize related work needed to understand research contributions leaving further details to [7]. The key terms for this research and frequently used expressions are introduced in Table I.

'Transfer-based' Others' Research: Methods [8], [4], [1] for IT service management applications propose generative models based on $P(transfer \mid ticket\ content)$ and recommend 'most probable' transfers between the experts. These models assume that the Markov property holds and thus recommend transfers only based on the previous node. (Note that this memoryless modeling does not considering workflow relationships between experts). In [5], a fixed short look-ahead subsequence was introduced to partially mitigate the problem caused by local modeling. However, all these transfer-based models used a common evaluation metric, i.e. MSTR, which *does not address the challenge of timely resolution mandated by the SLT Goal above*.

'TRS-workflow-based' Prior Research: We next differentiate our own previous stream of research as 'TRS-workflow-based' models using $P(TRS \mid ticket\ content)$, and briefly discuss our earlier results [7] used as the basis for the current paper's research. The closer examination of transfer sequences of resolving tickets revealed existence of 'collective' behaviors: 1) the skills of an expert apply based on previous experts and this might be repeated when resolving; and 2) sometimes longer transfer sequences do better in meeting SLTs. *Here the entire sequence of experts (which may include repeating nodes) is called the TRS*. We found that the longer TRSs arise from the complexity of virtual layers and multiple components. Thus making demands of specific expertise in the context of what has happened previously in the workflow. *We thus asked the question "can we improve performance w.r.t. time by considering TRS workflows globally rather than just local transfers?"*.

B. Current Research Motivation & Overview

Our previous research in [7] only partially answered the above question. Validation found that while resolution accuracy increases by recommending R-TRSs (workflows), the TTR estimates (ETTR) are deviated from actual TTRs (ATTR) resulting in high time estimation errors. This discrepancy warrants further research here for two reasons:

- A discrepancy between ETTR and ATTR for ticket t could be desirable! Particularly, if ETTR is less than ATTR, that could signal an improvement resulted from taking a RecTRS that is more efficient. This is significant because currently SLTs are relaxed to accommodate worst case scenarios. By providing methods to improve the resolution time for individual recommendations, we also help improve the SLT goal setting and manage resources.
- When deploying a recommendation system within IT operations, the trustworthiness of the RecTRSs should be achieved by precisely estimating TTRs. This results in trust from the professionals whose performance is measured by meeting time constraints.

To address this, Sections II and III present the integration of our previous and current research into a more complete 'training, testing and assessment' framework in Figure 1. By applying this framework we show 1) that for certain resolved tickets, their TTR is deviated from ETTR of their SLT-preserving RecTRS, and 2) that if this discrepancy better understood can provide opportunities for improving not only TTR, but also SLT in aggregate. To do this, the assessment method of Section III also analyzes the error of ETTRs and causes for estimation 'errors'; thus identifying those regions where the 'errors' indicate scope for further time improvement.

The next step is to understand the reaction of the experts to ticket content that is 'surprising' and consequent increases in resolution time. This is achieved by building a language model for each R-TRSs during training, and measuring the cross-entropy of a test ticket w.r.t. its RecTRS's language model. This allows us to verify whether high time estimation error is correlated with content that is deviated from the inherent language model of RecTRS.

Results in Section IV show that tickets with high cross-entropy or 'surprising' content are strongly correlated with the high ETTR errors. This actually means we need a better estimation model that captures not only the dynamics of surprise, but also all other factors contributing to the dynamic reaction of the experts leading to high estimation errors. Using this result as a basis we complete the framework details for selecting RecTRSs based on improved time estimates by considering the experts dynamics. Further research and conclusions are in Section V.

II. Analysis of CEN Achievement of SLT Goals

Through exploratory analysis it was found that 24% of all the TRSs have three or more experts (sometimes repeated more than once) to provide add-on and contributory knowledge. We identified this type of problem solving detected as **collective** problem solving [6]. We also use the more conceptual term below **'Collective Expert Network' or CEN** to refer to expert networks with this behavior. The CEN on a set of resolved tickets T is a directed graph where experts and transfers in T are represented by nodes and edges, respectively.

A. Enterprise Data Characteristics

The dataset for this research is obtained from the IT infrastructure and operations of a large insurance company using complex clouds and systems supporting best practices in IT service management. The rich ticket content included incident description content, transfer logs, timestamps related to transfers, actual time to resolve, time-sensitive service level alerts, etc. This dataset includes 150,000 tickets along with processing timestamps generated by over 900 technical support teams which include more than three thousand personnel. On average around 20,000 incidents are arisen from 7,500 configuration items each month. Also on average around 8,000 incidents are reported to the service desk by service clients.

SLT Related Definitions and Data Characterizations: For each type of ticket and service, the priority is negotiated according to business needs in collaboration with the customer. The priority sets a 'target' for time to resolve known as SLT. E.g. "Priority 1 tickets have to be resolved in 14 hours"; though, some tickets might be completed earlier. It is important to note that the SLT is a somewhat relaxed time constraint associated with the a range of incidents of a certain priority level. Thus the TTR may be greater or less than that SLT. Also the SLT is more relaxed for lower priorities. The SL clock runs per ticket, and all the experts along its TRS path contribute to the TTR. The SLT is said to be **'breached'** if the TTR is greater than the SLT. The priority level is set between P1 (highest impact) to P4 (negligible impact) based on the criticality to the customer and the type of problem solving needed. Our analysis found that tickets may have 1 to 18 associated transfers before resolution, and tickets with more transfers are more likely to breach their SLTs. We also found that the CEN tries to resolve as many tickets possible (85.4%) in one-to-two transfers after the service desk.

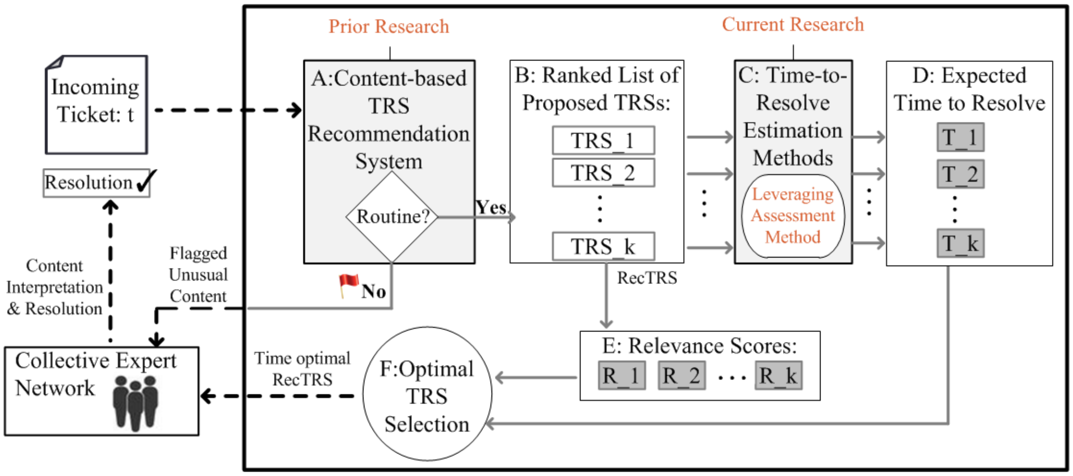
Prior Research
Current Research
Incoming Ticket: t
A:Content-based TRS Recommendation System
Routine?
Yes
No
B: Ranked List of Proposed TRSs:
TRS_1
TRS_2
TRS_k
C: Time-to-Resolve Estimation Methods
Leveraging Assessment Method
D: Expected Time to Resolve
T_1
T_2
T_k
RecTRS
E: Relevance Scores:
R_1
R_2
R_k
F:Optimal TRS Selection
Time optimal RecTRS
Collective Expert Network
Resolution ✓
Content Interpretation & Resolution
Flagged Unusual Content

Fig. 1: Overview of Time Optimal TRS Recommendation Framework

B. TRS-based Recommendation Results

In an enterprise with the service desk as the first node that sets the clock, our previous research [6] found that the CEN works hard at meeting SLTs and is more successful on **Routine** or frequent ticket content. We discovered an important pattern that frequent ticket content is 'highly relevant' to certain frequent TRSs. Furthermore, of those frequent TRSs approx. 98% met their associated SLTs. This provided the basis for the recommendation framework in Figure 1:

Recommendations based on content classification: Incoming ticket is classified using a two-level classification framework introduced by our prior research in [7]. The top level classifier labels the ticket content as Routine or Non-Routine (Figure 1 Box A). If the content is labeled as 'Non-Routine' then it is not used for further recommendations, but flagged for unassisted expert-driven resolution process. This helps retain only those tickets for which there is solid classification evidence ensuring greater accuracy to promote trust in the recommendations. If the ticket gets labeled as Routine, it will be followed by a second level classification which recommends a ranked list of TRSs based on the classification confidence for the incoming ticket (Figure 1 Box B).

Meeting the Resolution Goal: By recommending the Routine TRSs on frequent content, prior research established a 34% improvement in the accuracy of the recommendations when compared to the greedy baseline. In addition, it was found that the two-level TRS classification model has high precision (77%) when TRSs are recommended. Thus, establishing that RecTRS is an existing resolving sequence with a high likelihood to meet its SLT.

Meeting the SLT Goal: Next there are two factors related to evaluating a proposed TRS: 1) SLT and 2) ETTR. However, in prior research we limited ourselves to SLT evaluation and found that 99.8% of the tickets in the history that achieved their SLT are also expected to achieve their SLT after taking RecTRS (i.e. SLT Recall = 0.998). This firmly established that the TRS recommendations are reliable and meet SLTs.

Moving to current research, we wish to identify opportunities to **improve on, and not simply meet,** SLTs. Thus, also improving the aggregate SLT performance of CEN on T. This requires us to examine ETTR vs ATTR for RecTRSs. Existing time estimation models for ETTR are very approximate. The goal is to have better methods for time estimation for RecTRSs in Figure 1 Box C. Furthermore, since existing research does not well-address recommendation and validation against time-constraints, it thus became important to first conduct research into an error assessment framework for TTR estimation. This is reported next.

III. Understanding ETTR to Improve Resolution Time Estimation

We first show that by developing and using a method for assessment (leveraged in Figure 1 Box C) we can motivate the design of better features for time estimation for Box C which can then in turn be used for the selection of an SLT-optimal RecTRS (Circle F). This expands prior work by taking into consideration not only SLT achievement, but also the estimated time performance of recommendations validated over actual resolution time. For developing this assessment, a held-out test set of 3,560 tickets were used from which 1,636 tickets received recommended TRSs from box B in Figure 1 (the remaining 1,924 were flagged as Non-Routine). The performance of resulting recommended TRSs is assessed and summarized in Table II. Specific steps underlying this analysis are as follows:

- For the fraction of test tickets for which the TRSs are recommended, we used an expectation model to further estimate their TTR. We used $T_{P,R-TRS}$ to denote a subset of the training set that includes all tickets with priority P that were resolved by a particular routine TRS $R-TRS$. For a test ticket τ with priority P and recommended path $RecTRS$ the ETTR is estimated as the mean TTR of all tickets in $T_{P,RecTRS}$, formally:

$$\tau.ettr = \frac{1}{\mid T_{P,RecTRS} \mid} \sum_{t \in T_{P,RecTRS}} t.ttr \qquad (1)$$

- In order to compare ticket with different priorities within a unified scale we normalized all ETTR and ATTR values

TABLE II: Assessment of RecTRSs - ETTR vs ATTR

Assessment	ETTR?ATTR	Proposed	Actual	% test tickets
Investigate	1: $>$	SLT Met	SLT Met	65.9% [1078]
Ignore	2: $>$	Breached	SLT Met	0.2% [3]
Better	3: $<=$	SLT Met	SLT Met	31.8% [521]
Better	4: $<=$	SLT Met	Breached	1.9% [31]
Human	5: $<=$	Breached	Breached	0.2% [3]
Human	6: $>$	Breached	Breached	0.0% [0]

by their corresponding SLTs, thus generating NETTR and NATTR values. As a result of normalization if NETTR>1 then recommended TRS is estimated to breach its SLT, and if NATTR>1 then its actual SLT was breached according to the ground truth.

- For a test ticket τ we define estimation error (squared error) as: $(\tau.nettr - \tau.nattr)^2$.

The resulting six regions are next subject to causal analysis that leads to design of better estimation models. To achieve a deeper understanding, Table II presents SLT, ATTR and ETTR properties of tickets within each region and an assessment in the first column. Note the last column reports the probability (and frequency) distribution of test tickets over different regions.

- **Region 1:** While meeting SLTs, t.ETTR > t.ATTR. This needs to be investigated due to the fact that the higher ETTR estimates could be due to inaccurate (means based on history) estimation method. This motivates the further analysis and potentially considering the CEN's dynamic features in the next section, and thus designing improved methods for Box D of Figure 1. The output of this can then be more accurate, resulting in reliable SLT achieving recommendations that take less time.
- **Region 2:** Here the proposed RecTRSs are not appropriate for recommendation and thus not investigated further.
- **Region 3:** Here the proposed RecTRSs are actually improving the TTR and used as RecTRSs utilized in final selection circle F in Figure 1.
- **Region 4:** Here the proposed TRSs are actually benefiting the business contractually by avoiding breaches and used as RecTRSs in final selection circle F of Figure 1.
- **Region 5:** Not used for recommendation, flagged and sent directly to humans in Box A of Figure 1.
- **Region 6:** Not used for recommendation, flagged and sent directly to humans Box A of Figure 1.

Using the assessment of the ETTRs in Table II, our next goal is to further 'Investigate' TTR estimation methods for Box C of Figure 1 to gain insights and identify features that can yield more precise TTR estimation and thus improve the RecTRS selection process in Box D. This motivates the design of more rigorous ETTR models.

IV. Evidences of Dynamic CEN Behaviors

Note that the 'Investigation' of tickets in Region 1 (motivated above) requires investigation of external features for new estimation models which will improve the framework of Figure 1 Boxes C, D, and F by selecting from high-confidence RecTRSs using reliably low TTRs. Thus, the result is a new TRS recommendation model which proposes a pareto-optimal TRS that is characterized by the optimal combination of high recommendation *confidence* (i.e. $P(TRS|t.content)$) and low t.ETTR. The estimation model in Section III leverages the *Mean* TTR of the RecTRS for a given priority, and thus lacks explicit consideration of ticket content. We therefore ask: *Could this be a cause for inaccurate estimation?*

A. Content Deviation vs ETTR Error

Path-Priority Language Models: For each test ticket τ with priority P, and recommended path $RecTRS$, we aim to relate the language used in $\tau.content$ to the language of all tickets in the history (training data) which had priority P and got resolved by τ's $RecTRS$. The idea here is to measure how surprising the incoming content is to the $RecTRS$. Here we need a reliable model for the linguistic state of ($Priority$,TRS) pairs. Therefore, we define a *path-priority language model* for each ($Priority$,TRS) pair in the training set. This is constructed using Bigram language models with Katz back-off smoothing [3].

Cross Entropy of Content: Next for a test ticket τ, we quantify its content deviation w.r.t. its corresponding language model for (P,$RecTRS$), using cross-entropy computation:

$$H(\tau, LM_{(P,RecTRS)}) = -\frac{1}{N}\sum_{i=1}^{N} logP_{LM_{(P,RecTRS)}}(b_i) \tag{2}$$

Here there are N bigrams in τ, represented as b_is and $P_{LM_{(P,RecTRS)}}(b_i)$ is the probability of the bigram b_i under the language model for (P,$RecTRS$). A higher cross entropy for a ticket implies more deviation from the linguistic state of its RecTRS. For each test ticket τ we compared its min-max normalized cross entropy (NCE) against its time estimation squared error (SE). For training, we used the content of 41,800 natural language tickets to build 118 unique language models. Then 3,200 test tickets were carefully sampled for experimentation where each was ensured to receive accurate RecTRS (that is, matching its actual TRS). Analysis reveals insights:

- In the condition where there is a large estimation error for a ticket (SE>5), the normalized cross entropy also happens to be large. The correlation analysis for this case resulted in $R^2 = 0.5156$ which signifies strong positive correlation between time estimation error and normalized cross entropy. In other words, when the resolution time is mis-estimated by a large margin, ticket content is largely deviated from its RecTRSs' language models. With no conditions on the estimation error the normalized cross entropy is only weakly correlated with the estimation error. The positive correlations are shown in Figure 2 to illustrate the existence of a linear relationship between SE and NCE (regression line summarizes the relationship as $NCE = 0.0443SE + 0.2903$ with $R^2 = 0.1194$). This relationship is demonstrated more transparently on the

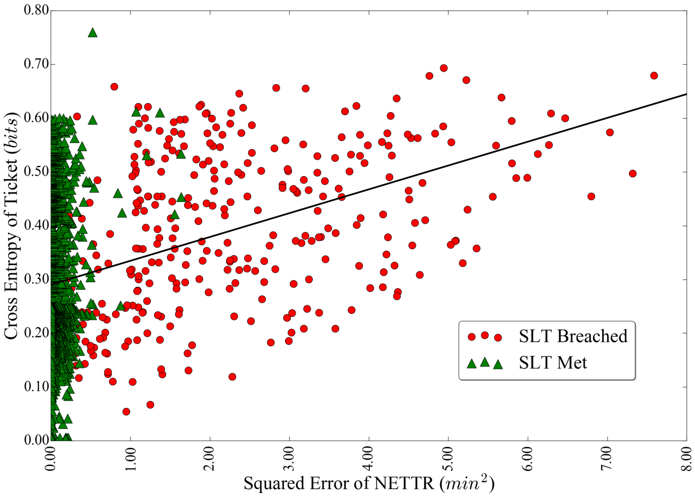

Fig. 2: Squared Error of NETTR vs Normalized Cross Entropy (per each test ticket)

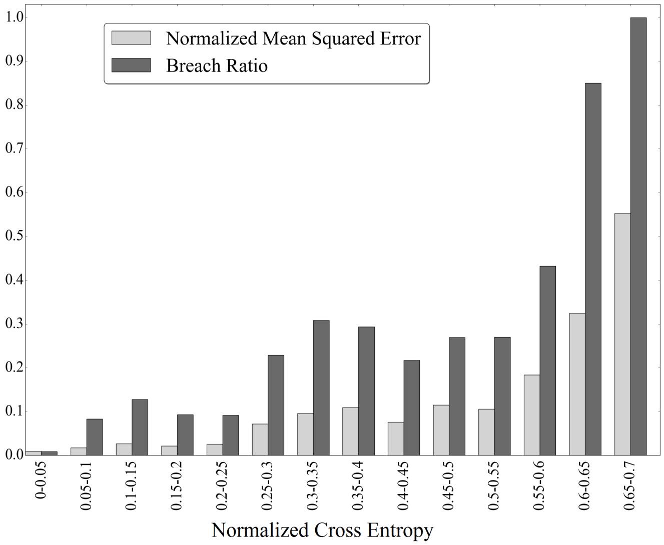

Fig. 3: Normalized Cross Entropy vs Breach Ratio and Normalized Mean Squared Error (aggregate level)

aggregate level in Figure 3, where larger NCE results in higher Normalized Mean Squared Error. However, the unconditioned relationship here is not as strong as the relationship where SE>5, due to the fact that a considerable fraction of tickets (26.4%) with low estimation error (SE<1) happen to have high normalized cross entropy (NCE>0.3). This indicates that **not** all tickets with high linguistic deviation are inherently complex for the CEN. This also means the linguistic models of historical TRSs alone cannot capture factors contributing to time estimation.

- High estimation errors mainly result from tickets with breached SLTs. 97.8% of tickets with SE>1 are breached (Figure 2). This uncovers a major pain point for path-based ETTR model in which 96.9% of tickets with an actual breached SLT will get estimated as meeting their SLTs. Therefore path-based ETTR model is incapable of (1) detecting such anomalous tickets, and (2) accurately estimating on them.
- SLT Breach ratio (likelihood to breach) increases as the NCE increases as shown in Figure 3. Thus, a dissimilarity metric between ticket content and the expertise of a TRS (such as NCE) qualifies as an important metric for better TTR estimation. This is in conformance with [10], which suggests that a breach in SLT generally happens due to unusual, complex or ambiguous content.

These insights lead us to requirements for a better TTR estimation model that must leverage dynamic features available early in the resolution process to detect anomalies (such as surprising content) and use them in the estimation process.

B. Further Research in CEN Dynamics

Motivated by the above analysis, our ongoing research goals are to achieve lower estimation error using an ensemble multivariate regression model defined at an expert-level. ETTR for a ticket is modeled as the sum of expected contribution time (ECT) by each intermediate TRS node, and expected resolution time (ERT) at the last node. Therefore dynamic time-related features that potentially affects each expert's response time are:

- *expertise profile at each node:* which is extracted by a novel iterative Expectation-Maximization algorithm which builds on logistic regression introduced in [9].
- *estimation of load at each node at time $\theta + \Delta\theta$:* which uses simulation grounded on the live load of the whole CEN at time θ. In a similar context, queue simulation has shown effectiveness in IT service engagement according to [2].

The next step is to develop underlying methods for inferring intrinsic features that affect the experts' response times. These features will feed into ECT/ERT models for each expert.

V. Conclusion

This research addresses a weakness in the treatment of time-related validation of recommendations. By assessing errors and causes, we show the need to go beyond error measures that use static historical data.The analysis shows that unusual content results in a ‘shock’ to the collective expert network perceptible as an increase in the SLT breach ratio, and an increase in the time estimation error. There remain other causes for poor estimation including network load. Further research in considering dynamic network features is thus motivated. The benefit of early and accurate resolution time estimation is not only *achievement* of SLTs, but also an *improvement* on the mean time to resolve of the tickets. Consequently, lower resolution times imply a decrease in average load of open tickets in the network, and an increase in experts' availability.

Acknowledgment

This research is supported by the National Science Foundation under Grant No. 0753710, the CERCS I/UCRC and industry sponsors.

References

[1] Yi Chen, Shu Tao, Xifeng Yan, Nikos Anerousis, and Qihong Shao. Assessing expertise awareness in resolution networks. In *Advances in Social Networks Analysis and Mining (ASONAM), International Conference on*, pages 128–135. IEEE, 2010.

[2] Yixin Diao, Linh Lam, Larisa Shwartz, and David Northcutt. Modeling the impact of service level agreements during service engagement. *IEEE Transactions on Network and Service Management*, 11(4):431–440, 2014.

[3] Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401, 1987.

[4] Gengxin Miao, Louise E Moser, Xifeng Yan, Shu Tao, Yi Chen, and Nikos Anerousis. Generative models for ticket resolution in expert networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 733–742. ACM, 2010.

[5] Gengxin Miao, Louise E Moser, Xifeng Yan, Shu Tao, Yi Chen, and Nikos Anerousis. Reliable ticket routing in expert networks. In *Reliable Knowledge Discovery*, pages 127–147. Springer, 2012.

[6] Kayhan Moharreri, Jayashree Ramanathan, and Rajiv Ramnath. Recommendations for achieving service levels within large-scale resolution service networks. In *Proceedings of the 8th Annual ACM India Conference*, pages 37–46. ACM, 2015.

[7] Kayhan Moharreri, Jayashree Ramanathan, and Rajiv Ramnath. Probabilistic sequence modeling for trustworthy it servicing by collective expert networks. In *Computer Software and Applications Conference (COMPSAC), IEEE 40th Annual*. IEEE, 2016.

[8] Qihong Shao, Yi Chen, Shu Tao, Xifeng Yan, and Nikos Anerousis. Efficient ticket routing by resolution sequence mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 605–613. ACM, 2008.

[9] Huan Sun, Mudhakar Srivatsa, Shulong Tan, Yang Li, Lance M Kaplan, Shu Tao, and Xifeng Yan. Analyzing expert behaviors in collaborative networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1486–1495. ACM, 2014.

[10] Wubai Zhou, Liang Tang, Tao Li, Larisa Shwartz, and Genady Ya Grabarnik. Resolution recommendation for event tickets in service management. In *IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 287–295. IEEE, 2015.