Appendix A: Perfect Recovery Guarantee for the Problem (5)

The following theorem shows the perfect recovery guarantee for the problem (5). Appendix C provides the proof for completeness.

**Theorem 7.1.** Let $X^* \in \mathbb{R}^{n \times n}$ be a rank $k$ matrix with a singular value decomposition $X^* = UV^T$, where $U = (u_1, \ldots, u_k) \in \mathbb{R}^{n \times k}$ and $V = (v_1, \ldots, v_k) \in \mathbb{R}^{n \times k}$ are the left and right singular vectors of $X^*$, respectively. Similar to many related works of matrix completion, we assume that the following two assumptions are satisfied:

1. The row and column spaces of $X$ have coherence bounded above by a positive number $\mu_0$.
2. Max absolute value in matrix $UV^T$ is bounded above by $\mu_1 \sqrt{r} / n$ for a positive number $\mu_1$.

Suppose that $m_1$ entries of $X^*$ are observed with their locations sampled uniformly at random, and among the $m_1$ observed entries, $m_2$ randomly sampled entries are corrupted. Using the resulting partially observed matrix as the input to the problem (5), then with a probability at least $1 - n^{-3}$, the underlying matrix $X^*$ can be perfectly recovered, given

1. $\mu(E) \xi(X) \leq \frac{1}{\log \tau}$,
2. $\frac{\xi(X) - (2k - 1) \mu(E) \xi^2(X)}{1 - 2(\log \tau) \mu(E) \xi(X)} < \lambda < \frac{1 - (4k + 5) \mu(E) \xi(X)}{(k + 2) \mu(E)}$, 
3. $m_1 - m_2 \geq C [\max(\mu_0, \mu_1)]^4 n \log^2 n,$

where $C$ is a positive constant; $\xi(\cdot)$ and $\mu(\cdot)$ denotes the low-rank and sparsity incoherence (Chandrasekaran et al., 2011).

Theorem 7.1 implies that even if some of the observed entries computed by (4) are incorrect, problem (5) can still perfectly recover the underlying similarity matrix $X^*$ if the number of observed correct entries is at least $O(n \log^2 n)$. For MATL with large $n$, this implies that only a tiny fraction of all task pairs is needed to reliably infer similarities over all task pairs. Moreover, the completed similarity matrix $X$ is symmetric, due to symmetry of the input matrix $Y$. This enables analysis by similarity-based clustering algorithms, such as spectral clustering.

Appendix B: Proof of Low-rankness of Matrix X

We first prove that the full similarity matrix $X \in \mathbb{R}^{n \times n}$ is of low-rank. To see this, let $A = (a_1, \ldots, a_k)$ be the underlying perfect clustering result, where $k$ is the number of clusters and $a_i \in \{0, 1\}^n$ is the membership vector for the $i$-th cluster. Given $A$, the similarity matrix $X$ is computed as

$$X = \sum_{i=1}^{k} a_i a_i^T = \sum_{i=1}^{k} B_i,$$

where $B_i = a_i a_i^T$ is a rank one matrix. Using the fact that $\text{rank}(X) \leq \sum_{i=1}^{k} \text{rank}(B_i)$ and $\text{rank}(B_i) = 1$, we have $\text{rank}(X) \leq k$, i.e., the rank of the similarity matrix $X$ is upper bounded by the number of clusters. Since the number of clusters is usually small, the similarity matrix $X$ should be of low rank.

Appendix C: Proof of Theorem 7.1

We then prove our main theorem. First, we define several notations that are used throughout the proof. Let $X = U\Sigma V^T$ be the singular value decomposition of matrix $X$, where $U = (u_1, \ldots, u_k) \in \mathbb{R}^{n \times k}$ and $V = (v_1, \ldots, v_k) \in \mathbb{R}^{n \times k}$ are the left and right singular vectors of matrix $X$, respectively. Similar to many related works of matrix completion, we assume that the following two assumptions are satisfied:

1. $A_1$: the row and column spaces of $X$ have coherence bounded above by a positive number $\mu_0$, i.e., $\sqrt{n/r} \max_i ||P_U(e_i)|| \leq \mu_0$ and $\sqrt{n/r} \max_i ||P_V(e_i)|| \leq \mu_0$, where $P_U = UU^T$, $P_V = VV^T$, and $e_i$ is the standard basis vector, and
2. **A2:** the matrix UV\(^{T}\) has a maximum entry bounded by \(\mu_{1}\sqrt{r}/n\) in absolute value for a positive number \(\mu_{1}\).

Let \(T\) be the space spanned by the elements of the form \(u_{i}y^{T}\) and \(xv^{T}\), for \(1 \leq i \leq k\), where \(x\) and \(y\) are arbitrary \(n\)-dimensional vectors. Let \(T^{\perp}\) be the orthogonal complement to the space \(T\), and let \(P_{T}\) be the orthogonal projection onto the subspace \(T\) given by

\[
P_{T}(Z) = P_{U}Z + ZP_{V} - P_{U}ZP_{V}.
\]

The following proposition shows that for any matrix \(Z \in T\), it is a zero matrix if enough amount of its entries are zero.

**Proposition 1.** Let \(\Omega\) be a set of \(m\) entries sampled uniformly at random from \([1, \ldots, n] \times [1, \ldots, n]\), and \(P_{\Omega}(Z)\) projects matrix \(Z\) onto the subset \(\Omega\). If \(m > m_{0}\), where \(m_{0} = C_{R}^{2}\mu_{0}rn\beta \log n\) with \(\beta > 1\) and \(C_{R}\) being a positive constant, then for any \(Z \in T\) with \(P_{\Omega}(Z) = 0\), we have \(Z = 0\) with probability \(1 - 3n^{-\beta}\).

**Proof.** According to the Theorem 3.2 in (Candès and Tao, 2010), for any \(Z \in T\), with a probability at least \(1 - 2n^{2-2\beta}\), we have

\[
\|P_{T}(Z)\|_{F} - \delta \|Z\|_{F} \leq \frac{n^{2}}{m} \|P_{T}P_{\Omega}(Z)\|_{F}^{2} = 0
\]

where \(\delta = m_{0}/m < 1\). Since \(Z \in T\), we have \(P_{T}(Z) = Z\). Then from (8), we have \(\|Z\|_{F} \leq 0\) and thus \(Z = 0\). \(\Box\)

In the following, we will develop a theorem for the dual certificate that guarantees the unique optimal solution to the following optimization problem

\[
\min_{X, E} \quad \|X\|_{*} + \lambda\|E\|_{1} \quad \text{s.t.} \quad P_{\Omega}(X + E) = P_{\Omega}(Y).
\]

**Theorem 1.** Suppose we observe \(m_{1}\) entries of \(X\) with locations sampled uniformly at random, denoted by \(\Delta\). We further assume that \(m_{2}\) entries randomly sampled from \(m_{1}\) observed entries are corrupted, denoted by \(\Delta\). Suppose that \(P_{\Omega}(Y) = P_{\Omega}(X+E)\) and the number of observed correct entries \(m_{1} - m_{2}\) \(= m_{0} = C_{R}^{2}\mu_{0}rn\beta \log n\). Then, for any \(\beta > 1\), with a probability at least \(1 - 3n^{-\beta}\), the underlying true matrices \((X, E)\) is the unique optimizer of (9) if both assumptions A1 and A2 are satisfied and there exists a dual \(Q \in \mathbb{R}^{n \times n}\) such that (a) \(Q = P_{\Omega}(Q)\), (b) \(P_{T}(Q) = UV^{T}\), (c) \(\|P_{T}(Q)\| < 1\), (d) \(P_{\Delta}(Q) = \lambda \mathrm{sgn}(E)\), and (e) \(\|P_{\Delta}(Q)\|_{\infty} < \lambda\).

**Proof.** First, the existence of \(Q\) satisfying the conditions (a) to (e) ensures that \((X, E)\) is an optimal solution. We only need to show its uniqueness and we prove it by contradiction. Assume there exists another optimal solution \((X + N_{X}, E + N_{E})\), where \(P_{\Omega}(N_{X} + N_{E}) = 0\). Then we have

\[
\|X + N_{X}\|_{*} + \lambda\|E + N_{E}\|_{1} \geq \|X\|_{*} + \lambda\|E\|_{1} + \langle Q_{E}, N_{E} \rangle + \langle Q_{X}, N_{X} \rangle
\]

where \(Q_{E}\) and \(Q_{X}\) satisfying \(P_{\Delta}(Q_{E}) = \lambda \mathrm{sgn}(E)\), \(\|P_{\Delta}(Q_{E})\|_{\infty} \leq \lambda\), \(P_{T}(Q_{X}) = UV^{T}\) and \(\|P_{T}(Q_{X})\| \leq 1\). As a result, we have

\[
\lambda\|E + N_{E}\|_{1} + \|X + N_{X}\|_{*} \\
\geq \lambda\|E\|_{1} + \|X\|_{*} + \langle Q + P_{\Delta}(Q_{E}) - P_{\Delta}(Q_{E}), N_{E} \rangle + \langle Q + P_{T}(Q_{X}) - P_{T}(Q_{X}), N_{X} \rangle \\
= \lambda\|E\|_{1} + \|X\|_{*} + \langle Q, N_{E} + N_{X} \rangle + \langle P_{T}(Q_{E}) - P_{\Delta}(Q_{E}), N_{E} \rangle + \langle P_{T}(Q_{X}) - P_{T}(Q_{X}), N_{X} \rangle \\
= \lambda\|E\|_{1} + \|X\|_{*} + \langle P_{\Delta}(Q_{E}) - P_{\Delta}(Q_{E}), P_{\Delta}(N_{E}) \rangle + \langle P_{T}(Q_{X}) - P_{T}(Q_{X}), P_{T}(Q_{X}) - P_{T}(Q_{X}), N_{X} \rangle
\]
We then choose $P_{\Delta^c}(Q_E)$ and $P_{T^\perp}(Q_X)$ to be such that $\langle P_{\Delta^c}(Q_E), P_{\Delta^c}(N_E) \rangle = \lambda \| P_{\Delta^c}(N_E) \|_1$ and $\langle P_{T^\perp}(Q_X), P_{T^\perp}(N_X) \rangle = \| P_{T^\perp}(N_X) \|_\ast$. We thus have

$$\lambda \| E + N_E \|_1 + \| X + N_X \|_\ast \geq \lambda \| E \|_1 + \| X \|_\ast + (\lambda - \| P_{\Delta^c}(Q) \|_\infty) \| P_{\Delta^c}(N_E) \|_1 + (1 - \| P_{T^\perp}(Q) \|) \| P_{T^\perp}(N_X) \|_\ast$$

Since $(X + N_X, E + N_E)$ is also an optimal solution, we have $\| P_{\Omega^c}(N_E) \|_1 = \| P_{T^\perp}(N_X) \|_\ast$, leading to $P_{\Omega^c}(N_E) = P_{T^\perp}(N_X) = 0$, or $N_X \in T$. Since $P_{\Omega}(N_X + N_E) = 0$, we have $N_X = N_E + Z$, where $P_{\Omega}(Z) = 0$ and $P_{\Omega}(N_X) = 0$. Hence, $P_{\Omega^c \cap \Omega}(N_X) = 0$, where $|\Omega^c \cap \Omega| = m_1 - m_2$. Since $m_1 - m_2 > m_0$, according to Proposition 1, we have, with a probability $1 - 3n^{-\beta}$, $N_X = 0$. Besides, since $P_{\Omega}(N_X + N_E) = P_{\Omega}(N_E) = 0$ and $\Delta \subset \Omega$, we have $P_{\Delta}(N_E) = 0$. Since $N_E = P_{\Delta}(N_E) + P_{\Delta^c}(N_E)$, we have $N_E = 0$, which leads to the contradiction.

Given Theorem 1, we are now ready to prove Theorem 3.1.

**Proof.** The key to the proof is to construct the matrix $Q$ that satisfies the conditions (a)-(e) specified in Theorem 1. First, according to Theorem 1, when $m_1 - m_2 > m_0 = C_0^2 \mu_0 r n \beta \log n$, with a probability at least $1 - 3n^{-\beta}$, mapping $P_T P_{\Omega} P_T(Z) : T \mapsto T$ is an one to one mapping and therefore its inverse mapping, denoted by $(P_T P_{\Omega} P_T)^{-1}$ is well defined. Similar to the proof of Theorem 2 in (Chandrasekaran et al., 2011), we construct the dual certificate $Q$ as follows

$$Q = \lambda \text{sgn}(E) + \epsilon_\Delta + P_{\Delta}(P_T P_{\Omega} P_T)^{-1}(UV^\top + \epsilon_T)$$

where $\epsilon_T \in T$ and $\epsilon_\Delta = P_{\Delta}(\epsilon_D)$. We further define

$$H = P_{\Omega} P_T (P_T P_{\Omega} P_T)^{-1}(UV^\top)$$

$$F = P_{\Omega} P_T (P_T P_{\Omega} P_T)^{-1}(\epsilon_T)$$

Evidently, we have $P_{\Omega}(Q) = Q$ since $\Delta \subset \Omega$, and therefore the condition (a) is satisfied. To satisfy the conditions (b)-(e), we need

1. $P_T(Q) = UV^\top \rightarrow \epsilon_T = -P_T(\lambda \text{sgn}(E) + \epsilon_\Delta)$
2. $\| P_{T^\perp}(Q) \| < 1 \rightarrow \mu(E)(\lambda + \| \epsilon_\Delta \|_\infty) + \| P_{T^\perp}(H) \| + \| P_{T^\perp}(F) \| < 1$
3. $P_{\Delta}(Q) = \lambda \text{sgn}(E) \rightarrow \epsilon_\Delta = -P_{\Delta}(H + F)$
4. $\| P_{\Delta^c}(Q) \|_\infty < \lambda \rightarrow \xi(X)(1 + \| \epsilon_T \|) < \lambda$

Below, we will first show that there exist solutions $\epsilon_T \in T$ and $\epsilon_\Delta$ that satisfy conditions (10) and (12). We will then bound $\| \epsilon_\Delta \|_\infty, \| \epsilon_T \|, \| P_{T^\perp}(H) \|$, and $\| P_{T^\perp}(F) \|$ to show that with sufficiently small $\mu(E)$ and $\xi(X)$, and appropriately chosen $\lambda$, conditions (11) and (13) can be satisfied as well.

First, we show the existence of $\epsilon_\Delta$ and $\epsilon_T$ that obey the relationships in (10) and (12). It is equivalent to show that there exists $\epsilon_T$ that satisfies the following relation

$$\epsilon_T = -P_T(\lambda \text{sgn}(E)) + P_T P_{\Delta}(H) + P_T P_{\Delta}(P_T P_{\Omega} P_T)^{-1}(\epsilon_T)$$

or

$$P_T P_{\Omega \setminus \Delta}(P_T P_{\Omega} P_T)^{-1}(\epsilon_T) = -P_T(\lambda \text{sgn}(E)) + P_T P_{\Delta}(H),$$

where $\Omega \setminus \Delta$ indicates the complement set of set $\Delta$ in $\Omega$ and $|\Omega \setminus \Delta|$ denotes its cardinality. Similar to the previous argument, when $|\Omega \setminus \Delta| = m_1 - m_2 > m_0$, with a probability $1 - 3n^{-\beta}$, $P_T P_{\Omega \setminus \Delta}(P_T(Z) : T \mapsto T$ is an one to one mapping, and therefore $(P_T P_{\Omega \setminus \Delta}(P_T(Z)))^{-1}$ is well defined. Using this result, we have the following solution to the above equation

$$\epsilon_T = P_T P_{\Omega}(P_T P_{\Omega \setminus \Delta} P_T)^{-1}(-P_T(\lambda \text{sgn}(E)) + P_T P_{\Delta}(H))$$
We now bound $\|\epsilon_T\|$ and $\|\epsilon_\Delta\|_\infty$. Since $\|\epsilon_T\| \leq \|\epsilon_T\|_F$, we bound $\|\epsilon_T\|_F$ instead. First, according to Corollary 3.5 in (Candès and Tao, 2010), when $\beta = 4$, with a probability $1 - n^{-3}$, for any $Z \in T$, we have

$$\|P_T \perp P_T (P_T \perp P_T)^{-1}(Z)\|_F \leq \|Z\|_F.$$  

Using this result, we have

$$\|\epsilon_\Delta\|_\infty \leq \xi(X)(\|H\| + \|F\|)$$

$$\leq \xi(X)(1 + \|P_T \perp (H)\|_F + \|\epsilon_T\|_F + \|P_T \perp (F)\|_F)$$

$$\leq \xi(X)(2 + \|\epsilon_T\| + \|\epsilon_T\|_F)$$

$$\leq \xi(X)[2 + (2k + 1)\|\epsilon_T\|]$$

In the last step, we use the fact that rank($\epsilon_T$) $\leq 2k$ if $\epsilon_T \in T$. We then proceed to bound $\|\epsilon_T\|$ as follows

$$\|\epsilon_T\| \leq \mu(E)(\lambda + \|\epsilon_\Delta\|_\infty)$$

Combining the above two inequalities together, we have

$$\|\epsilon_T\| \leq \xi(X)\mu(E)(2k + 1)\|\epsilon_T\| + 2\xi(X)\mu(E) + \lambda\mu(E)$$

$$\|\epsilon_\Delta\|_\infty \leq \xi(X)[2 + (2k + 1)\mu(E)(\lambda + \|\epsilon_\Delta\|_\infty)]$$

which lead to

$$\|\epsilon_T\| \leq \frac{\lambda\mu(E) + 2\xi(X)\mu(E)}{1 - (2k + 1)\xi(X)\mu(E)}$$

$$\|\epsilon_\Delta\|_\infty \leq \frac{2\xi(X) + (2k + 1)\lambda\xi(X)\mu(E)}{1 - (2k + 1)\xi(X)\mu(E)}$$

Using the bound for $\|\epsilon_\Delta\|_\infty$ and $\|\epsilon_T\|$, we now check the condition (11)

$$1 > \mu(E)(\lambda + \|\epsilon_\Delta\|_\infty) + \frac{1}{2} + \frac{k}{2}\|\epsilon_T\|$$

or

$$\lambda < \frac{1 - \xi(X)\mu(E)(4k + 5)}{\mu(E)(k + 2)}$$

For the condition (13), we have

$$\lambda > \xi(X)\mu(E)(\|\epsilon_T\|)$$

or

$$\lambda > \frac{\xi(X) - (2k + 1)\xi^2(X)\mu(E)}{1 - 2k + 1\xi(X)\mu(E)}$$

To ensure that there exists $\lambda \geq 0$ satisfies the above two conditions, we have

$$1 - 5(2k + 1)\xi(X)\mu(E) + (10k^2 + 21k + 8)[\xi(X)\mu(E)]^2 > 0$$

and

$$1 - \xi(X)\mu(E)(4k + 5) \geq 0$$

Since the first condition is guaranteed to be satisfied for $k \geq 1$, we have

$$\xi(X)\mu(E) \leq \frac{1}{4k + 5}.$$  

Thus we finish the proof.

\textbf{Appendix D: Data Statistics}

We listed the detailed domains of the sentiment analysis tasks in Table 3. We removed the \textit{musical instruments} and \textit{tools hardware} domains from the original data because they have too few labeled examples. The statistics for the 10 target tasks of intent classification in Table 4.
<table>
<thead>
<tr>
<th>Domains</th>
<th>#train</th>
<th>#validation</th>
<th>#test</th>
</tr>
</thead>
<tbody>
<tr>
<td>apparel</td>
<td>7398</td>
<td>926</td>
<td>928</td>
</tr>
<tr>
<td>automotive</td>
<td>601</td>
<td>69</td>
<td>66</td>
</tr>
<tr>
<td>baby</td>
<td>3405</td>
<td>437</td>
<td>414</td>
</tr>
<tr>
<td>beauty</td>
<td>2305</td>
<td>280</td>
<td>299</td>
</tr>
<tr>
<td>books</td>
<td>19913</td>
<td>2436</td>
<td>2489</td>
</tr>
<tr>
<td>camera_photo</td>
<td>5915</td>
<td>744</td>
<td>749</td>
</tr>
<tr>
<td>cell_phones_service</td>
<td>816</td>
<td>109</td>
<td>98</td>
</tr>
<tr>
<td>computer_video_games</td>
<td>2201</td>
<td>274</td>
<td>296</td>
</tr>
<tr>
<td>dvd</td>
<td>19961</td>
<td>2624</td>
<td>2412</td>
</tr>
<tr>
<td>electronics</td>
<td>18431</td>
<td>2304</td>
<td>2274</td>
</tr>
<tr>
<td>gourmet_food</td>
<td>1227</td>
<td>182</td>
<td>166</td>
</tr>
<tr>
<td>grocery</td>
<td>2101</td>
<td>268</td>
<td>263</td>
</tr>
<tr>
<td>health_personal_care</td>
<td>5826</td>
<td>687</td>
<td>712</td>
</tr>
<tr>
<td>jewelry_watches</td>
<td>1597</td>
<td>188</td>
<td>196</td>
</tr>
<tr>
<td>kitchen_housewares</td>
<td>15888</td>
<td>1978</td>
<td>1990</td>
</tr>
<tr>
<td>magazines</td>
<td>3341</td>
<td>427</td>
<td>421</td>
</tr>
<tr>
<td>music</td>
<td>20103</td>
<td>2463</td>
<td>2510</td>
</tr>
<tr>
<td>office_products</td>
<td>337</td>
<td>54</td>
<td>40</td>
</tr>
<tr>
<td>outdoor_living</td>
<td>1321</td>
<td>143</td>
<td>135</td>
</tr>
<tr>
<td>software</td>
<td>1934</td>
<td>254</td>
<td>202</td>
</tr>
<tr>
<td>sports_outdoors</td>
<td>4582</td>
<td>566</td>
<td>580</td>
</tr>
<tr>
<td>toys_games</td>
<td>10634</td>
<td>1267</td>
<td>1246</td>
</tr>
<tr>
<td>video</td>
<td>19941</td>
<td>2519</td>
<td>2539</td>
</tr>
</tbody>
</table>

Table 3: Statistics of the Multi-Domain Sentiment Classification Data.

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>#labeled instances</th>
<th>#labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>497</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>3071</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>305</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>122</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>110</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>126</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>218</td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>297</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>424</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>110</td>
<td>17</td>
</tr>
</tbody>
</table>

Table 4: Statistics of the User Intent Classification Data.