



Measurement of the Top Pair Production Cross Section using RomaNN in the Lepton Plus Jets Decay Channel with 4.3 fb^{-1}

The CDF Collaboration
URL <http://www-cdf.fnal.gov>
(Dated: January 15, 2010)

The cross-section for pair produced top quarks in the lepton plus jets channel has been measured in 4.3 fb^{-1} of collected data from the high p_t lepton triggers. To improve signal significance, the measurement identifies jets produced by bottom quark decays, using a bottom “tagging” algorithm called RomaNN. Events are required to have at least one “tagged” jet. The result is $\sigma_{t\bar{t}} = 6.88 \pm 0.29_{stat} \pm 0.83_{sys} \pm 0.42_{lumi}$.

I. INTRODUCTION

We present a measurement of the top pair cross-section using $4.3fb^{-1}$ of collected data from the CDF detector [1]. Data is selected using an inclusive high Pt lepton trigger requiring an electron or muon with at least 20 GeV. In addition, we require missing transverse energy $\cancel{E}_T > 20$ GeV, at least three jets present in the event with $E_t > 20$ GeV, and the scalar sum of the transverse energy (Ht) of the jets, lepton, and \cancel{E}_T to be greater than 230 GeV. A “tagging” algorithm, RomaNN, in its “Tight” operating point, is used to identify jets with bottom quark decay.

In general, the cross section is calculated with the formula:

$$\sigma_{t\bar{t}} = \frac{N_{data} - N_{bkg}}{A \cdot \epsilon \cdot L} \quad (1)$$

where, N_{data} is the amount of collected data in the signal region, N_{bkg} is the predicted background content, A is the acceptance of $t\bar{t}$ events before requiring the jet to be identified as coming from a bottom quark decay (b-tagging), ϵ is the b-tag efficiency, and L is the luminosity.

Monte Carlo simulations are relied upon to estimate acceptance and tagging efficiency, though with corrections applied to account for mismodeling in trigger efficiencies, lepton identification efficiency, b-tagging efficiency, and mistag probability. We determine the value of the cross section as the one that maximizes the likelihood for the data to be consistent with the predicted background plus the top signal as a function of its cross section. Systematic uncertainties are calculated by varying the parameters one by one under $\pm 1\sigma$ deviations and re-performing the measurement.

II. ROMANN “TAGGING” ALGORITHM

The identification of b-jets is an essential component for measurements in the top quark sector and searches for a low mass Higgs boson and other new phenomena. The signatures of these interesting signal processes all contain b jets; the ability to discriminate b jets from the overwhelming inclusive jet background helps increase the purity of the selected event sample.

At CDF, so far b-jets have been identified using an algorithm capable to reconstruct secondary vertices, and placing requirements on the significance of this displaced vertex. Recently, a new tagging algorithm has been developed, called RomaNN, and is different from the standard tagging algorithms in that it is a multivariate tool, incorporating information from several sources simultaneously. It incorporates information from track impact parameters, a possible secondary vertex, semileptonic decays, as well as other variables in its per-jet tagging decision.

RomaNN provides a per-jet output value in the range from -1 to +1 therefore each analysis has an opportunity to customize the level of b purity by choosing its own cut in the RomaNN output value. This analysis uses the “Tight” operating point.

III. BACKGROUND ESTIMATE

To model the background it’s not possible to rely only on MC simulation, since there are inadequacies in the description of the production rate for heavy flavor in association with a W boson, of the tagging efficiency for bottom jets, and difficulties associated with modeling the QCD contribution. So, we take a data-driven approach that is combined with the Monte Carlo simulations, using a technique that is sequential, where each step depends on the previous.

The technique relies on the lepton plus jets sample before the jets are required to have come from a b-quark decays (pre-tagged sample), to determine the overall normalization of the processes, and then predict the content of the lepton plus b-tagged jets sample (tagged sample) by estimating the tagging efficiency for the different processes using simulations. The final result is a complete prediction for the process content in the lepton plus jets data sample. In the following we will go step by step through the procedure.

A. Electroweak Backgrounds

A few of the backgrounds which are considered a small contribution to the overall process content and $t\bar{t}$ (which is an important point as we will discuss later) are estimated relying on the Monte Carlo simulation. Several electroweak

processes contribute to the lepton plus jets sample such as WW, WZ, ZZ, and $Z \rightarrow jets$ events. They exist in the sample because each process can produce a real lepton and neutrino, as well as a number of jets.

The yields of events from these processes in the final sample are estimated using the theoretical cross section, the luminosity of the sample, trigger efficiency, and an overall selection efficiency derived from Monte Carlo simulation of the processes. The calculated number in our sample is given by

$$N_{ewk}^{pretag} = \sigma_{p\bar{p} \rightarrow X} \cdot A \cdot \int dt \cdot \mathcal{L} \quad (2)$$

$$N_{ewk}^{tag} = \sigma_{p\bar{p} \rightarrow X} \cdot A \cdot \epsilon \cdot \int dt \cdot \mathcal{L} \quad (3)$$

where $\sigma_{p\bar{p} \rightarrow X}$ is the theoretical cross sections, $\int dt \cdot \mathcal{L}$ is the total luminosity, A is the pre-tagged selection acceptance derived from Monte Carlo, and ϵ is the tagged selection efficiency. The top signal estimate is estimated in the same manner as the electroweak backgrounds.

B. Non-W Based Background Estimate

Part of the background in the lepton + jets sample come from QCD events, where one jet is mis-identified as a lepton and some \cancel{E}_T is created in the events if the jets energy is not correctly measured. To generate these rare events in MC it would be a highly time consuming task, and it would be difficult to simulate them correctly, so this background is described using data itself. To do so, we use the lepton plus jets sample, where some of the identification requirements on the lepton has been reversed.

To estimate the overall normalization of the this background, we fit the \cancel{E}_T distribution of the non- W template and the MC template for the other backgrounds to data.

Both data and model templates are fitted to the \cancel{E}_T distribution of isolated pretag data events using a binned likelihood fit. Once the fraction is calculated the normalization is simply:

$$N_{QCD}^{pretag} = F_{QCD} \cdot N_{pretag} \quad (4)$$

The same general procedure is performed for the tagged sample.

$$N_{QCD}^{tag} = F_{QCD} \cdot N_{tag} \quad (5)$$

C. W + Heavy Flavor

W plus jets is the catch-all category for events that are not considered QCD, electroweak, or top. In the pretag data sample, the W plus jets normalization is calculated by subtracting the electroweak processes and the QCD from data as shown in equation 6.

$$N_{W+Jets}^{pretag} = N_{pretag} \cdot (1 - F_{QCD}^{pretag}) - N_{ewk}^{pretag} - N_{t\bar{t}}^{pretag} \quad (6)$$

For the tagged estimate, the W plus jets sample is broken down into two categories: heavy and light flavor, these two processes produce a tagged jet very differently and therefore requires different treatment in calculating the normalization.

The contribution of the heavy flavor background to our signal region is calculated by equation 7.

$$N_{W+hf}^{tag} = (N_{pretag} \cdot (1 - F_{QCD}) - N_{ewk}^{pretag} - N_{t\bar{t}}^{pretag}) \cdot f_{HF} \cdot K \cdot \epsilon \quad (7)$$

where f_{HF} is the fraction of events with jets matched to heavy flavor quarks, K is a correction to the Monte Carlo heavy flavor fraction called the ‘‘K-factor’’, and ϵ is the tagging efficiency.

The f_{HF} is calculated from a detailed Monte Carlo simulation Alpgen [2], and includes all possible processes contributing to the production of a single real W-boson.

The f_{HF} and ϵ are calculated for $Wb\bar{b}$, $Wc\bar{c}$, and Wc separately, which define the rates for each of these processes. Only the heavy flavor fraction relies on Monte Carlo, the normalization is derived from the pretag sample in data. The HF correction is derived by a Neural Network fit to variables sensitive to jets matched to heavy flavor and light flavor.

D. Mistags

A light flavor jet that is misidentified as a b-jet is called a mistag. The mistag rate for the RomaNN is handled using a mistag matrix. A small complication arises because there is no concept of symmetry in the RomaNN, therefore the concept of using a negative tag rate is not available for the RomaNN, and we have to measure the overall tag rate and subtract from it the tag rate due to heavy flavor.

This technique is applied to estimate the number of events in our sample due to mistags in W + light flavor events. The predicted number of background events from W + light flavor (W+lf) processes is:

$$N_{W+lf}^{tag} = \frac{N^{mistag}}{N^{pretag}} \cdot (N^{pretag} - N_{t\bar{t}}^{pretag} - N_{QCD}^{pretag} - N_{W+h_f}^{pretag} - N_{ewk}^{pretag}) \quad (8)$$

Where N^{mistag} is the predicted number of mistags in the event. The predicted amount of $t\bar{t}$, and QCD, W+hf, Electroweak background events is subtracted from the total pretag sample leaving an estimate for the W+lf fraction. The predicted number of mistagged W+lf events is the W+lf fraction multiplied by the predicted amount of mis-tagged events from the pretag data.

E. Full Background Prediction

Table I shows the background estimate used in our top pair production cross section measurement utilizing 4.3 fb^{-1} of collected data.

Process	1jet	2jets	3jets	4jets	5jets
Pretag Data	6411.0 ± 0.0	7785.0 ± 0.0	4617.0 ± 0.0	2080.0 ± 0.0	633.0 ± 0.0
Top (7.4pb)	10.3 ± 1.6	175.0 ± 25.3	557.6 ± 80.0	644.1 ± 91.3	221.9 ± 32.0
WW	3.4 ± 1.2	18.5 ± 4.2	13.0 ± 2.8	5.3 ± 1.2	1.8 ± 0.4
WZ	1.0 ± 0.2	5.3 ± 0.8	3.9 ± 0.6	1.7 ± 0.3	0.5 ± 0.1
ZZ	0.1 ± 0.0	0.6 ± 0.1	0.8 ± 0.1	0.4 ± 0.1	0.1 ± 0.0
Stop S	1.3 ± 0.2	26.9 ± 3.0	15.4 ± 1.7	4.4 ± 0.5	1.0 ± 0.1
Stop T	0.4 ± 0.1	25.4 ± 2.9	15.3 ± 1.6	4.3 ± 0.4	0.7 ± 0.1
Z+jets	3.6 ± 1.9	12.2 ± 3.3	11.7 ± 2.5	4.8 ± 1.0	1.4 ± 0.3
Wbb	50.0 ± 16.0	139.7 ± 44.3	101.2 ± 32.4	40.5 ± 13.7	12.6 ± 4.6
Wcc	34.9 ± 12.3	80.6 ± 27.6	64.3 ± 22.0	27.1 ± 9.7	8.7 ± 3.3
Wcj	31.5 ± 11.1	54.6 ± 18.7	28.3 ± 9.7	9.0 ± 3.2	2.3 ± 0.9
Mistags	81.6 ± 64.9	132.1 ± 48.3	76.6 ± 25.2	26.1 ± 10.5	7.1 ± 3.8
Non-W	47.6 ± 15.3	104.5 ± 31.8	61.8 ± 18.5	18.8 ± 16.0	6.9 ± 6.5
Total Prediction	265.7 ± 77.2	775.3 ± 110.9	949.9 ± 110.3	786.4 ± 98.5	265.0 ± 34.4
Observed	267.0 ± 0.0	716.0 ± 0.0	876.0 ± 0.0	760.0 ± 0.0	281.0 ± 0.0

TABLE I: Predicted and observed for ≥ 1 Tag, $HT \geq 230$ GeV, and $\cancel{E}_T \geq 20$ GeV

IV. CALCULATING THE CROSS-SECTION

With the background estimate in hand it would appear straightforward to calculate the cross section, but because the background estimate is dependent on the top pair production cross section, extracting the measured value is not so simple. To do that we construct a poisson likelihood where the background's dependence on the signal estimate is taken into account. The likelihood is:

$$-2 \cdot \ln L = -2 \cdot (N_{data} \cdot \ln(D \cdot \sigma_{t\bar{t}} + B(\sigma_{t\bar{t}})) - \ln(N_{data}!) - (D \cdot \sigma_{t\bar{t}} + B(\sigma_{t\bar{t}}))) \quad (9)$$

where $D = A \cdot \epsilon \cdot L$ is the denominator of equation 1, N_{data} is the amount of measured data, and $B(\sigma_{t\bar{t}})$ is the background estimate for a given top pair production cross section. The likelihood is calculated for several values of the cross section and the resulting points are fit to a second order polynomial. The minimum of this curve is taken as the measured value. The result for our selection, $H_T \geq 230$ GeV and $\cancel{E}_T \geq 20$ GeV, is $\sigma_{t\bar{t}} = 6.88 \pm 0.29_{stat}$ pb, and the fit is shown in Figure 1.

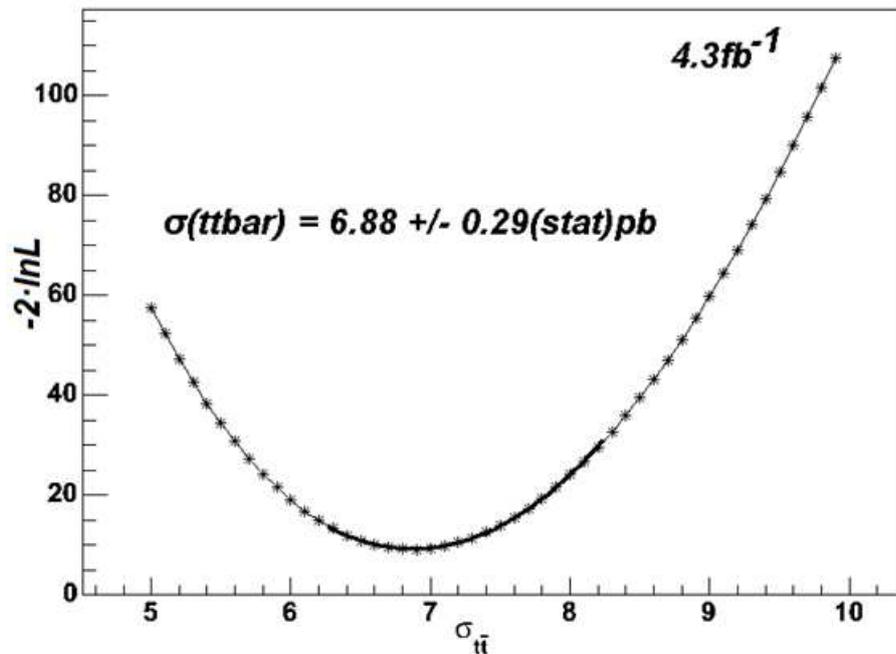


FIG. 1: Likelihood Curve For Measured Cross Section

V. SYSTEMATICS

Systematic uncertainties in our measurement result are calculated by varying a given parameter within its uncertainty and redoing the entire measurement. Each systematic is described below along with any relevant quantities. The individual evaluated systematic uncertainties are shown in Table II.

Systematic	$\Delta\sigma$	$\Delta\sigma/\sigma$
Luminosity	0.42	6.0%
K Factor	0.23	3.3%
B Tag SF	0.59	8.5%
C Tag SF	0.16	2.3%
Mistag Matrix	0.26	3.8%
QCD Fraction	0.10	1.5%
Color Recon	0.06	0.8%
JES	0.36	5.2%
ISR/FSR	0.16	2.3%
MC Generator	0.18	2.6%
CEM SF	0.01	0.1%
CMUP SF	0.01	0.1%
CMX SF	0.01	0.1%
PDF	0.02	0.3%
Total	0.94	13.5%

TABLE II: Systematic Uncertainties

A. Luminosity

The uncertainty on our calculated luminosity is derived from the CLC accuracy and the uncertainty on the theoretical cross section of inelastic $p\bar{p}$ collisions. The luminosity is fluctuated within this uncertainty and the measurement redone.

B. K-Factor

The correction to the heavy flavor fractions has an uncertainty derived from the Neural Network fits in the 1 and 2 jet bin as well as the fits to bottom and charm separately. The K-factor is varied by its 20% uncertainty and the measurement redone.

C. Tagging

Because MC does not model b-tagging properly, a scale factor is applied to each tagged jet matched to heavy flavor, and the corresponding event then re-weighted. The scale factor is derived from data and has an uncertainty associated with it which leads to a systematic on the measurement. The effect on the measured value is calculated by fluctuating the scale factor within its uncertainty, applying it to each appropriate jet, calculating the new event weights, and repeating the measurement.

D. Mistag Matrix

Mistags are not reliably modeled in simulations so we use a data-driven parameterization called the mistag matrix to predict the probability that any given jet is mistagged. The mistag rate on any jet fluctuated by 40% up(down) and the entire measurement is repeated to quantify the effect.

E. QCD Fractions

To estimate the uncertainty on the QCD fraction, the fits are shifted by 30% and the measurement repeated, the resulting difference in the result is taken as a systematic uncertainty in the measurement.

F. Color Reconnection

To study this effect, we replace our standard $t\bar{t}$ Monte Carlo model with two different tunes Apro and ACRpro and the measurement is redone taking the absolute difference.

G. Jet Energy Scale

The energy of jets measured by the calorimeters is subject to multiple systematic uncertainties. We study the effect on the measurement by varying the JES for our top signal Monte Carlo and background models and then re-performing the measurement. The effect of JES on this measurement is mainly through the acceptance of signal and background.

H. Initial/Final State Radiation

The measured value will be effected if we are over or under estimating the amount of initial or final state radiation present in top events. To study this effect, we replace our standard top Monte Carlo model with two top Monte Carlos where the radiation has been increased/decreased and the measurement is redone.

I. Parton Shower Modeling

Differences in Monte Carlo shower models are studied simply by replacing our $t\bar{t}$ PYTHIA model with the other most popular generator, HERWIG, and repeating the measurement [3] [4]

J. Trigger Efficiency

Detector specific corrections are applied to the Monte Carlo to more correctly model the relative trigger efficiencies between CEM, CMUP, and CMX events. The corrections are data-derived from Z events and have a small uncertainty associated with them. There are two types of corrections, trigger ID and trigger efficiencies. Each are fluctuated with their uncertainty, separately, and the resulting errors are added in quadrature.

K. PDF

Uncertainty in the parton distribution function are evaluated by a re-weighting scheme at the Monte Carlo Truth level. PDF's are reweighted in our signal Monte Carlo to simulate 46 different PDF parameterizations, and the measurement is performed for each different parameterization.

VI. RESULT

Extracting the result from the likelihood and adding the systematic uncertainty we find the cross section of $4.3fb^{-1}$ using ≥ 1 Tight RomaNN Tagged events in the lepton plus jets channel is:

$$\sigma_{t\bar{t}} = 6.88 \pm 0.29_{\text{stat}} \pm 0.83_{\text{sys}} \pm 0.42_{\text{lum}} \text{ pb} \quad (10)$$

Acknowledgments

We thank the Fermilab staff and the technical staffs of the participating institutions for their vital contributions. This work was supported by the U.S. Department of Energy and National Science Foundation; the Italian Istituto Nazionale di Fisica Nucleare; the Ministry of Education, Culture, Sports, Science and Technology of Japan; the Natural Sciences and Engineering Research Council of Canada; the National Science Council of the Republic of China; the Swiss National Science Foundation; the A.P. Sloan Foundation; the Bundesministerium für Bildung und Forschung, Germany; the Korean Science and Engineering Foundation and the Korean Research Foundation; the Science and Technology Facilities Council and the Royal Society, UK; the Institut National de Physique Nucleaire et Physique des Particules/CNRS; the Russian Foundation for Basic Research; the Comisión Interministerial de Ciencia y Tecnología, Spain; the European Community's Human Potential Programme; the Slovak R&D Agency; and the Academy of Finland.

-
- [1] F. Abe et al., Nucl. Instrum. Methods Phys. Res A 271, 387 (1988)
D. Amidei, et al, Nucl. Instrum. Methods Phys. Res. A 350, 73 (1994)
F. Abe et al., Phys Rev D 52, 4784 (1995)
P. Azzi et al., Nucl. Instrum. Methods Phys. Res A 360, 137(1995)
CDFII Technical Design Report, FERMILAB-PUB-96/390-E
 - [2] M.L. Mangano, M. Moretti, F. Piccinini, R. Pittau, A. Polosa, "ALPGEN, a generator for hard multiparton processes in hadronic collisions", JHEP 0307:001,2003, hep-ph/0206293.
 - [3] T.Sjostrand et al., Comput. Phys. Commun. 135, 238, (2001)
 - [4] G.Corcella et al., JHEP 01,10 (2001)