CDF Note 9445

Search for Single Top Quark Production using Boosted Decision Trees in 2.7 fb^{-1} of CDF Data

The CDF Collaboration
URL http://www-cdf.fnal.gov
(Dated: July 25, 2008)

We present a search for electroweak single top quark production using 2.7 fb^{-1} of CDF II data collected between February 2002 and April 2008 at the Tevatron in proton-antiproton collisions at a center-of-mass energy of 1.96 TeV. The analysis employes a multivariate technique based on Boosted Decision Trees, where the output is used to build a discriminant variable which we will fit to the data using a binned likelihood approach. We search for a combined single top s- and t-channel signal and measure a cross section of $2.4^{+0.8}_{-0.7}$ pb assuming a top quark mass of 175 GeV/c^2. The probability that the observed excess originated from a background fluctuation (p-value) is 0.00017 (3.6σ) and the expected (median) p-value in pseudo-experiments is 4.1×10^{-7} which corresponds to a 4.9σ signal significance assuming single top quark production at the rate predicted by the Standard Model.

Preliminary Results for the Summer Conferences 2008

INTRODUCTION

In proton anti-proton collisions at the Tevatron with a center-of-mass energy of 1.96 TeV, top quarks are predominantly produced in pairs via the strong force. In addition, the Standard Model predicts single top quarks to be produced through an electroweak t- and s-channel exchange of a virtual W boson as shown in Figure 1. The production cross sections have been calculated at Next-to-Leading-Order (NLO). For a top quark mass of 175 GeV/c^2 the results are 1.98±0.25 pb and 0.884±0.11 pb for the t-channel and s-channel process respectively [1]. The combined cross section is about 40% of the top anti-top pair production cross section ($\sigma_{singletop} \sim 2.9$ pb). The measurement of electroweak single top production probes the $W - t - b$ vertex, provides a direct determination of the Cabbibo-Kobayashi-Maskawa (CKM) matrix element $|V_{tb}|$ and offers a source of almost 100% polarized top quarks [2]. Moreover, the search for single top also probes exotic models beyond the Standard Model. New physics, like flavor-changing neutral currents or heavy W' bosons, could alter the observed production rate [3]. Finally, single top processes result in the same final state as the Standard Model Higgs boson process $WH \to Wb\bar{b}$, which is one of the most promising low mass Higgs search channels at the Tevatron [4]. Essentially, all analysis tools developed for the single top search can be used for this Higgs search.

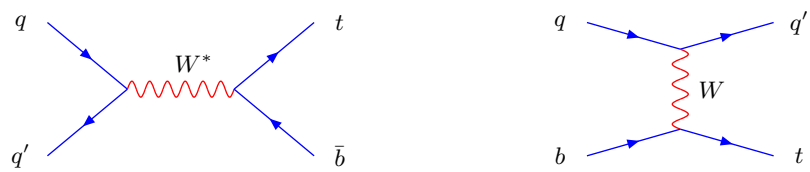

FIG. 1: Leading order Feynman diagrams for s-channel (left) and t-channel (right) single top quark production.

Finding single top quark production is challenging since it is rarely produced in comparison with other processes with the same final state like W+jets and $t\bar{t}$. The signal to background ratio of the analysis is small, typically on the order of less than S/B∼1/15. This calls for a better discrimination of signal and background events which can be achieved by using more information to characterize each event. We have employed a new analysis approach at CDF that attempts to make optimal use of information in the data by means of a multivariate technique via Boosted Decision Trees (BDT).

DATA SAMPLE & EVENT SELECTION

Our single top event selection exploits the kinematic features of the signal final state, which contains a real W boson, one or two bottom quarks, and possibly additional jets. To reduce multi-jet backgrounds, the W originating from the top quark decay is required to have decayed leptonically. We demand therefore a high-energy electron or muon ($E_T(e) > 20$ GeV, or $P_T(\mu) > 20$ GeV/c) and large missing transverse energy (MET) from the undetected neutrino MET>25 GeV. Electrons are measured in the central and in the forward calorimeter, $|\eta| < 1.6$. Exactly two or three jets with $E_T > 20$ GeV and $|\eta| < 2.8$ are required to be present in the event. A large fraction of the backgrounds is removed by demanding at least one of these two jets to be tagged as a b-quark jet by using displaced secondary vertex information from the silicon vertex detector. The secondary vertex tagging algorithm identifies tracks associated with the jet originating from a vertex displaced from the primary vertex indicative of decay particles from relatively long lived B mesons. The backgrounds surviving these selections are $t\bar{t}$, W + heavy-flavor jets, i.e. $W + b\bar{b}$, $W + c\bar{c}$, $W + c$ and diboson events WW, WZ, and ZZ. Instrumental backgrounds originate from mis-tagged W + jets events (W events with light-flavor jets, i.e. with u, d, s-quark and gluon content, misidentified as heavy-flavor jets) and from non-W + jets events (multi-jet events where one jet is erroneously identified as a lepton).

BACKGROUND ESTIMATE

Estimating the background contribution after applying the event selection to the single top candidate sample is an elaborate process. NLO cross section calculations exist for diboson and $t\bar{t}$ production, thereby making the estimation of their contribution a relatively straightforward process. The main background contributions are from $W + b\bar{b}$, $W + c\bar{c}$ and $W + c$ + jets, as well as mis-tagged W + light quark jets. We determine the W + jets normalization from the data

and estimate the fraction of the candidate events with heavy-flavor jets using ALPGEN Monte Carlo samples [5]. The heavy-flavor fractions were calibrated in the b-tagged $W + 1$ jet sample using data distributions which are sensitive to distinguish light-flavor from heavy-flavor jets, e.g. the mass of the secondary-vertex and, more sophisticated, the output of the Neural Network jet-flavor separator. Based on these studies, the heavy flavor content was corrected by a factor $K_{HF} = 1.4 \pm 0.4$. The probability that a W + light-flavor jet is mis-tagged is parameterized using large statistics generic multi-jet data. The instrumental background contribution from non-W events is estimated using side-band data with low missing transverse energy, devoid of any signal, and we subsequently extrapolate the contribution into the signal region with large missing transverse energy, MET>25 GeV. The expected signal and background yield in the $W + 2$ jet and $W + 3$ jet sample is shown in Table I and graphically as a function of $W +$ jet multiplicity next to the table. These yields contain an additional acceptance, with respect to previous single top analyses at CDF, by including an extra muon coverege from events triggered via a MET + 2 jets trigger, which are complementary to the inclusive high p_T-lepton triggers used in previous single top CDF analyses. The new muon coverage, shown in Fig. 2, increases the muon signal acceptance by about 30% while keeping a smaller increase in background acceptance since the jet requirments at trigger level are more efficient accepting signal-like events.

Process	Number of Events in 2.7 fb^{-1}	
	W + 2 jets	W + 3 jets
s-channel	49.3 ± 7.0	16.3 ± 2.3
t-channel	74.3 ± 10.9	22.3 ± 3.2
$Wb\bar{b}$	549.1 ± 165.5	169.8 ± 51.3
$Wc\bar{c}, Wcj$	453.5 ± 139.6	126.7 ± 39.0
$Mistags$	410.7 ± 51.0	125.5 ± 15.8
$non-W$	75.6 ± 30.2	27.4 ± 11.0
WW	49.4 ± 5.5	18.0 ± 2.0
WZ	24.4 ± 1.9	7.3 ± 0.6
ZZ	0.7 ± 0.1	0.3 ± 0.0
$Z + jets$	31.1 ± 4.6	13.4 ± 2.0
$t\bar{t}$ dilepton	58.8 ± 8.4	51.3 ± 7.3
$t\bar{t}$ non-dilepton	114.7 ± 16.4	359.2 ± 51.1
Total signal	123.6 ± 13.0	38.6 ± 3.9
Total prediction	1891.1 ± 312.4	937.5 ± 108.3
Observed in data	1874	902

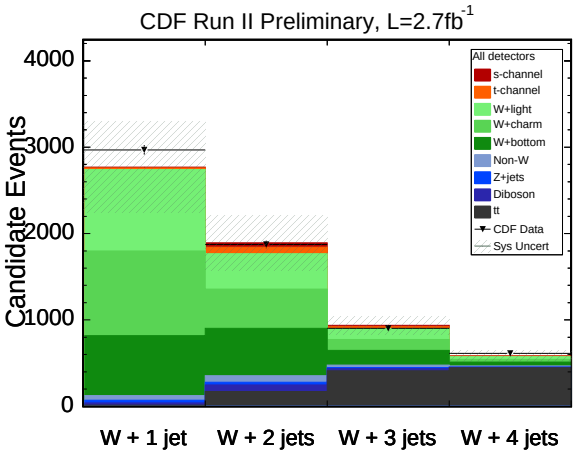

TABLE I: Number of expected single top and background events in 2.7 fb^{-1} of CDF II data passing all event selection cuts (left). Graphical representation of the predicted and observed W+jets yield (right).

ANALYSIS METHOD

In order to search for a single top quark production we developed a multivariate technique based on Boosted Decision Trees. To Build the BDTs we make use of the ROOT-integrated package TMVA [7].

A decision tree is a binary tree structured classifier like the one sketched in Fig. 3. Repeated left/right (yes/no) decisions are performed on a single variable at a time until some stop criterion is reached. Like this the phase space is split into regions that are eventually classified as signal or background, depending on which makes up the majority of training events that end up in the final leaf nodes. The boosting of a decision tree (BDT) represents an extension to a single decision tree. Several decision trees (a forest), derived from the same training sample by reweighting events, are combined to form a classifier which is given by a (weighted) majority vote of the individual decision trees. This process, called boosting, stabilizes the response of the decision trees with respect to fluctuations in the training sample.

Using boosted decision trees many kinematic or event shape input variables are combined into a single output variable with powerful discriminantion between signal and background. In the search for single top quark production, four different boosted decision trees are trained in different jet and b-tag bins:

- 2 jets, 1 b-tag
- 2 jets, 2 b-tags
- 3 jets, 1 b-tag

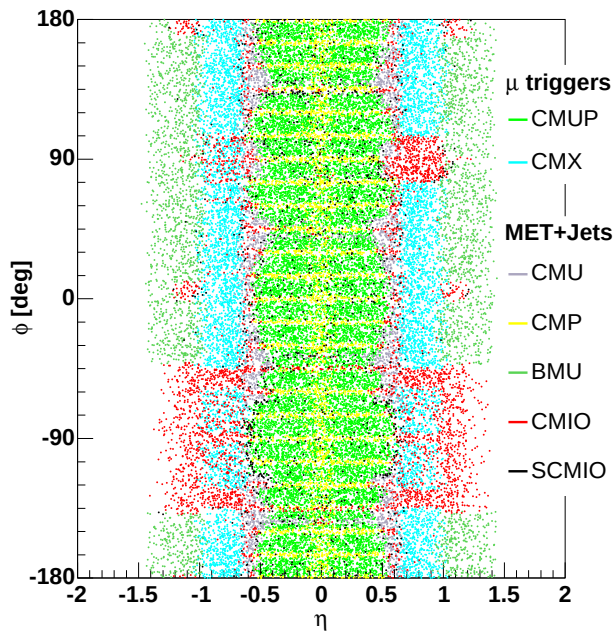

FIG. 2: This plot shows the increased acceptance of muon + jets events triggered through MET + 2jets trigger in addition to muon + jets events triggered through the default inclusive high p_T-muon trigger.

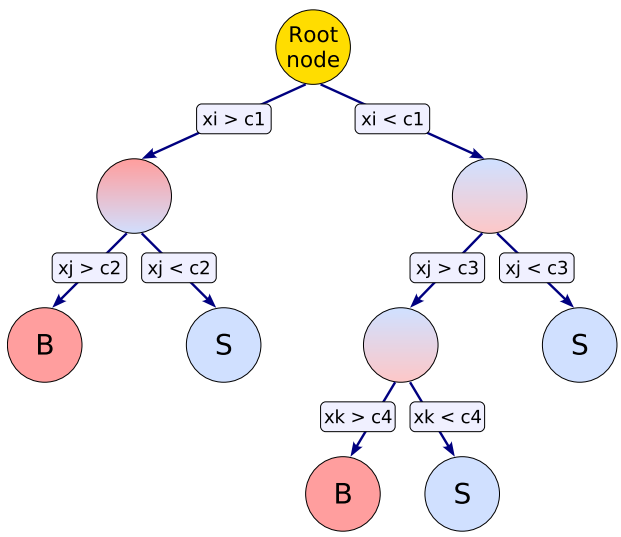

FIG. 3: Schematic view of a decision tree. Starting from the root node, a sequence of binary splits using the discriminating variables x_i is performed. Each split uses the variable that at this node gives the best separation between signal and background when being cut on. The same variable may thus be used at several nodes, while others might not be used at all. The leaf nodes at the bottom end of the tree are labeled S for signal and B for background depending on the majority of events that end up in the respective nodes.

- 3 jets, $\geq$ 2 b-tags

The level of agreement between data and Monte Carlo Simulation is checked for all the input variables in the four signal regions as well as in different control regions. In addition to the validation of the input variables, the output of the four trained BDTs are found to be in good agreement with data in three different control regions: 2 and 3 jets with no b-tags (W + light flavor dominant) and 4 jets with at least 1 b-tag ($t\bar{t}$ dominant).

We construct template histograms for signal and background. The templates for all signal and background processes for the BDT optimized in the four signal regions are shown in Fig. 4.

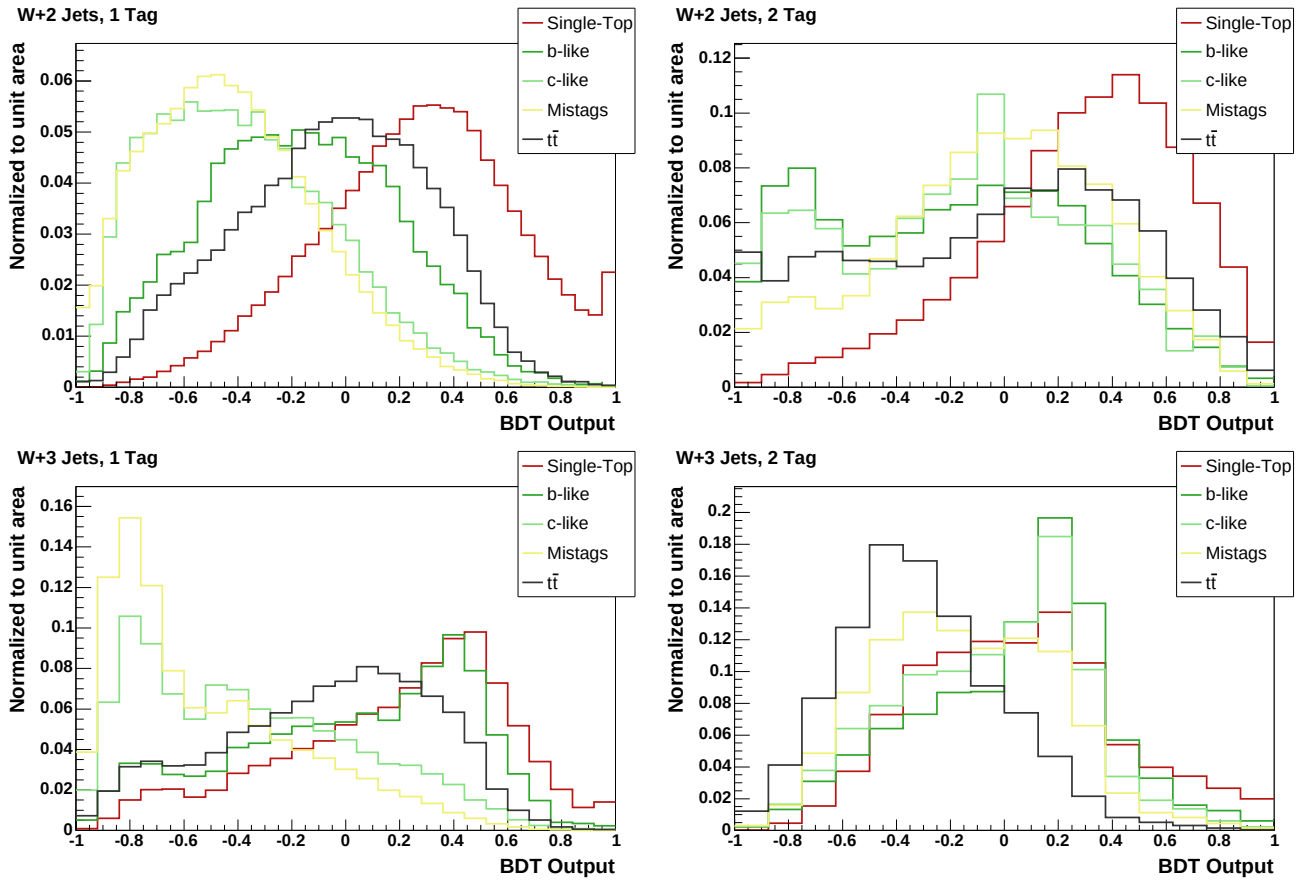

FIG. 4: Templates of the BDT outputs. The top two plots show the two-jet bin while the bottom plots show the three-jet bin. Single-tag discriminants are on the left side, while double-tag discriminants are on the right side. All histograms are normalized to unit area.

We perform a binned maximum likelihood fit to the data, in which the background templates are Gaussian constrained (within their respective uncertainties) to the predicted background yield while the signal template is free floating in the fit. The likelihood fit result determines the most probable value of the single-top cross section. Sources of systematic uncertainty are accounted for in the definition of the likelihood function shown in Equation 1.

SYSTEMATIC UNCERTAINTIES

We address systematic uncertainty from several different sources: (1) jet energy scale (2), initial state radiation (3), final state radiation , (4) parton distribution functions, (5) the event generator, (6) the uncertainty in the event detection efficiency, (7) the uncertainty on the integrated luminosity, (8) neural network b-tagger uncertainty, (9) ALPGEN Monte Carlo factorization/renormalization scale uncertainty, (10) uncertainty on the mistag model, (11) uncertainty on the non-W model, and (12) uncertainty on the Monte Carlo modeling. Systematic uncertainties can influence both, the expected event yield (normalization) and the shape of the discriminant distribution.

Normalization uncertainties are estimated by calculating the variation in the expected event yield due to a systematic effect. The range of systematic rate and shape variations across signal and background processes are shown in Table II. Shape uncertainties are estimated by producing shifted template histograms for each process due to the systematic effect. The bin-by-bin relative variations are used as shape systematics in the likelihood function. The letter 'X' in Table II indicates that a shape systematic has been evaluated for the particular nuisance parameter and included in the likelihood function.

For all backgrounds the normalization uncertainties are represented by the uncertainty on the predicted number of background events and are incorporated in the analysis as Gaussian constraints $G(\beta_j|1,\Delta_j)$ in the likelihood function:

Systematic Uncertainty	Rate Uncertainty	Shape Uncertainty
Jet energy scale	0...16%	X
Initial state radiation	0...11%	X
Final state radiation	0...15%	X
Parton distribution functions	2...3%	X
Monte Carlo generator	1...5%	
Event detection efficiency	0...9%	
Luminosity	6%	
Neural Network Jet-Flavor Separator	N/A	X
Fact. Ren. Scale in Alpgen MC	N/A	X
Mistag model	N/A	X
non-W	N/A	X
MC mis-modeling	N/A	X
W+bottom normalization	30%	
W+charm normalization	30%	
Mistag normalization	17...29%	
$t\bar{t}$ normalization	23%	

TABLE II: Minimum to maximum range of observed systematic normalization variations estiamted across all different processes and analysis input channels. The X indicates that a template shape uncertainty has been evaluated for that particular nuisance parameter and has been included in the likelihood function.

$$\mathcal{L}(\beta_1,\dots,\beta_5;\delta_1,\dots,\delta_{10}) = \underbrace{\prod_{k=1}^{B}\frac{e^{-\mu_k}\cdot\mu_k^{n_k}}{n_k!}}_{Poisson\ term}\cdot\underbrace{\prod_{j=2}^{5}G(\beta_j|1,\Delta_j)}_{Gauss\ constraints}\cdot\underbrace{\prod_{i=1}^{12}G(\delta_i,0,1)}_{Systematics} \tag{1}$$

$$where,\ \mu_k = \sum_{j=1}^{5}\beta_j\ \cdot\ \underbrace{\left\{\prod_{i=1}^{12}\left[1+|\delta_i|\cdot(\epsilon_{ji+}H(\delta_i)+\epsilon_{ji-}H(-\delta_i))\right]\right\}}_{Normalization\ Uncertainty} \tag{2}$$

$$\underbrace{\cdot\ \alpha_{jk}}_{Shape\ P.}\ \cdot\ \underbrace{\left\{\prod_{i=1}^{12}\left(1+|\delta_i|\cdot(\kappa_{jik+}H(\delta_i)+\kappa_{jik-}H(-\delta_i))\right)\right\}}_{Shape\ Uncertainty} \tag{3}$$

The systematic normalization and shape uncertainties are incorporated into the likelihood as nuisance parameters, conforming with a fully Bayesian treatment [6]. We take the correlation between normalization and shape uncertainties for a given source into account [8]. The relative strength of a systematic effect due to the source i is parameterized by the nuisance parameter δ_i in the likelihood function, constrained to a unit-width Gaussian (last term in Equation 1). The $\pm 1\sigma$ changes in the normalization of process j due to the ith source of systematic uncertainty are denoted by $\epsilon_{ji}+$ and $\epsilon_{ji}-$ (see Equation part 2). The $\pm 1\sigma$ changes in bin k of the EPD templates for process j due to the ith source of systematic uncertainty are quantified by κ_{jik+} and κ_{jik-} (see Equation part 3). $H(\delta_i)$ represents the Heaviside function, defined as $H(\delta_i) = 1$ for $\delta_i > 0$ and $H(\delta_i) = 0$ for $\delta_i < 0$. The Heaviside function is used to separate positive and negative systematic shifts (for which we have different normalization and shape uncertainties). The variable δ_i appears in both the term for the normalization (Equation 2) and the shape uncertainty (Equation 3), which is how correlations between both effects are taken into account. We reduce the likelihood function to the parameter of interest (the single top cross-section) by the standard Bayesian marginalizing procedure [9].

RESULTS

We apply the analysis to 2.7 $fb^{-1}$of CDF Run II Data. In order to extract the most probable single top content in the data we perform a maximum likelihood fit of the event probability discriminant distributions. The posterior

p.d.f is obtained by using Bayes' theorem:

$$p(\beta_1|data) = \frac{\mathcal{L}^*(data|\beta_1)\pi(\beta_1)}{\int \mathcal{L}^*(data|\beta_1')\pi(\beta_1')d\beta_1'}$$

where $\mathcal{L}^*(data|\beta_1)$ is the marginalized likelihood and $\pi(\beta_1)$ is the prior p.d.f. for β_1. We adopt a flat prior, $\pi(\beta_1) = H(\beta_1)$, in this analysis, with H being the Heaviside step function.

The most probable value corresponds to the most likely single top production cross section given the data. The uncertainty corresponds to the range of highest posterior probability density which covers 68.27%. Performing the likelihood fit with all systematic rate and shape uncertainties included in the likelihood function, we measure a single top cross section of $2.4^{+0.8}_{-0.7}$ pb. The posterior probability density is shown on the left of Fig. 5. In the Fig. 6 the BDT output distribution for signal and background, normalized to the Standard Model prediction, are shown in the four signal regions and compared to the distributions in data.

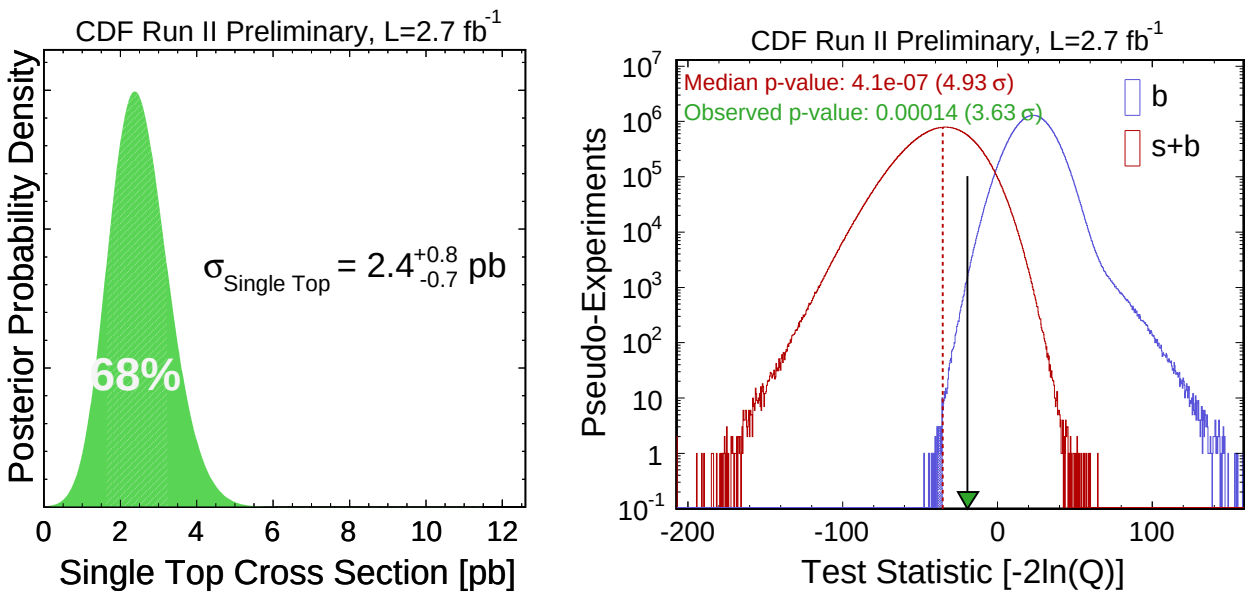

FIG. 5: Left: cross section result using 2.7 $fb^{-1}$of CDF II data. The error band shows the 68% uncertainty (all systematics included) on the measurement. Right: Distribution of the likelihood ratio test statistic for the signal + background (s+b) and background only hypothesis. The arrow indicates the result observed in data and the red dashed line indicates the expected median result.

We have calculated the signal significance of this result using a standard likelihood ratio technique [10]. In this approach, pseudo-experiments are generated from background only events. We define the likelihood ratio test statistic $Q = -2\ln\frac{P(data)|(s+b)}{P(data)|(b)}$ and calculate the p-value, i.e. the probability of the background only hypothesis (b) to fluctuate to the observed result in data or higher. We estimate the expected p-value, by taking the median of the test hypothesis (s+b) distribution as the 'observed' value (dashed red line in right plot of Fig. 5). We expect a p-value of 4.1×10^{-7} (4.9 σ) and observe a p-value of 0.00017 (3.6 σ) in the data. All sources of systematic uncertainty are included in our statistical treatment and we consider correlation between normalization and discriminant shape changes due to sources of systematic uncertainty (e.g. the jet-energy-scale uncertainty) as described in the previous section.

SINGLE TOP SIGNAL FEATURES

In Fig.7 we enrich our candidate sample with single top events by making increasing cuts on our BDT output and look for single top signal features for a few sensitive variables like $Q_{lepton} \cdot \eta_{untagged\ jet}$ which shows a distinct asymmetry for t-channel single top events and the invariant mass of the $W - b$ system, a quantity which should be close to the top quark mass. Although the uncertainties are large, there is a good shape agreement between data and the Monte Carlo prediction including single top (all plots normalized to the observed data).

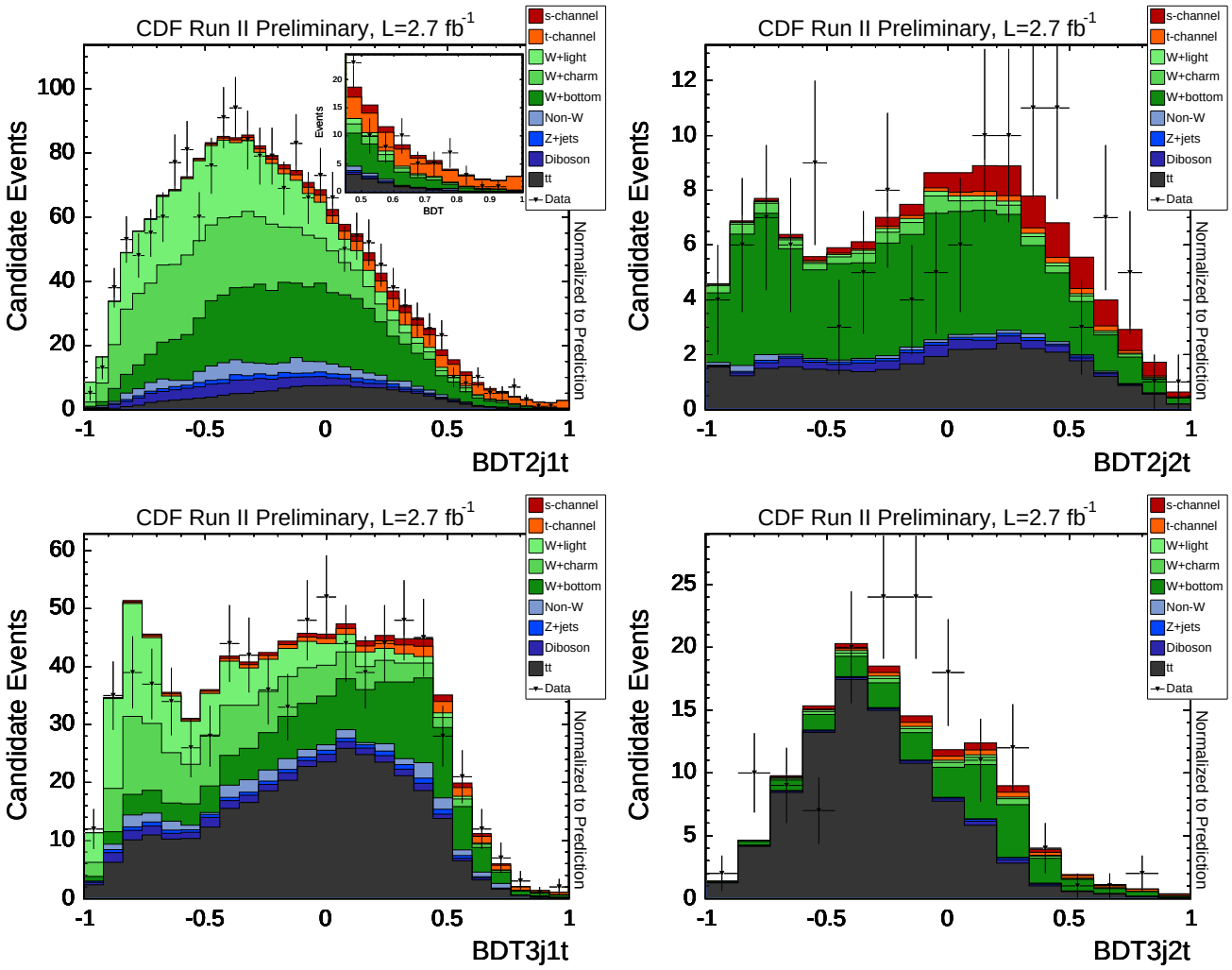

FIG. 6: CDF data compared to Monte Carlo prediction for signal and background. The top plots show the two-jet bin while the bottom plot shows the three-jet bin; the left-hand plots show the single-tagged events while the right-hand plots show the double-tagged events.

CONCLUSIONS

We report a measurement of electroweak single top quark production at CDF II using 2.7 $fb^{-1}$of proton-antiproton collisions recorded at the Tevatron. We employ a new multivariate technique at CDF based on Boosted Decision Trees for this search and measure a combined s-channel and t-channel single top cross-section of $\sigma_{singletop} = 2.4^{+0.8}_{-0.7}$ pb assuming a top quark mass of 175 GeV/c^2. We use a standard likelihood ratio technique to calculate the signal significance. The observed p-value is 0.00017 (3.6 σ) and the expected (median) p-value in pseudo-experiments is 4.1×10^{-7} (4.9 σ).

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 Ministerio de Educación y Ciencia and Programa Consolider-Ingenio 2010, Spain; the Slovak R&D Agency; and the Academy of Finland.

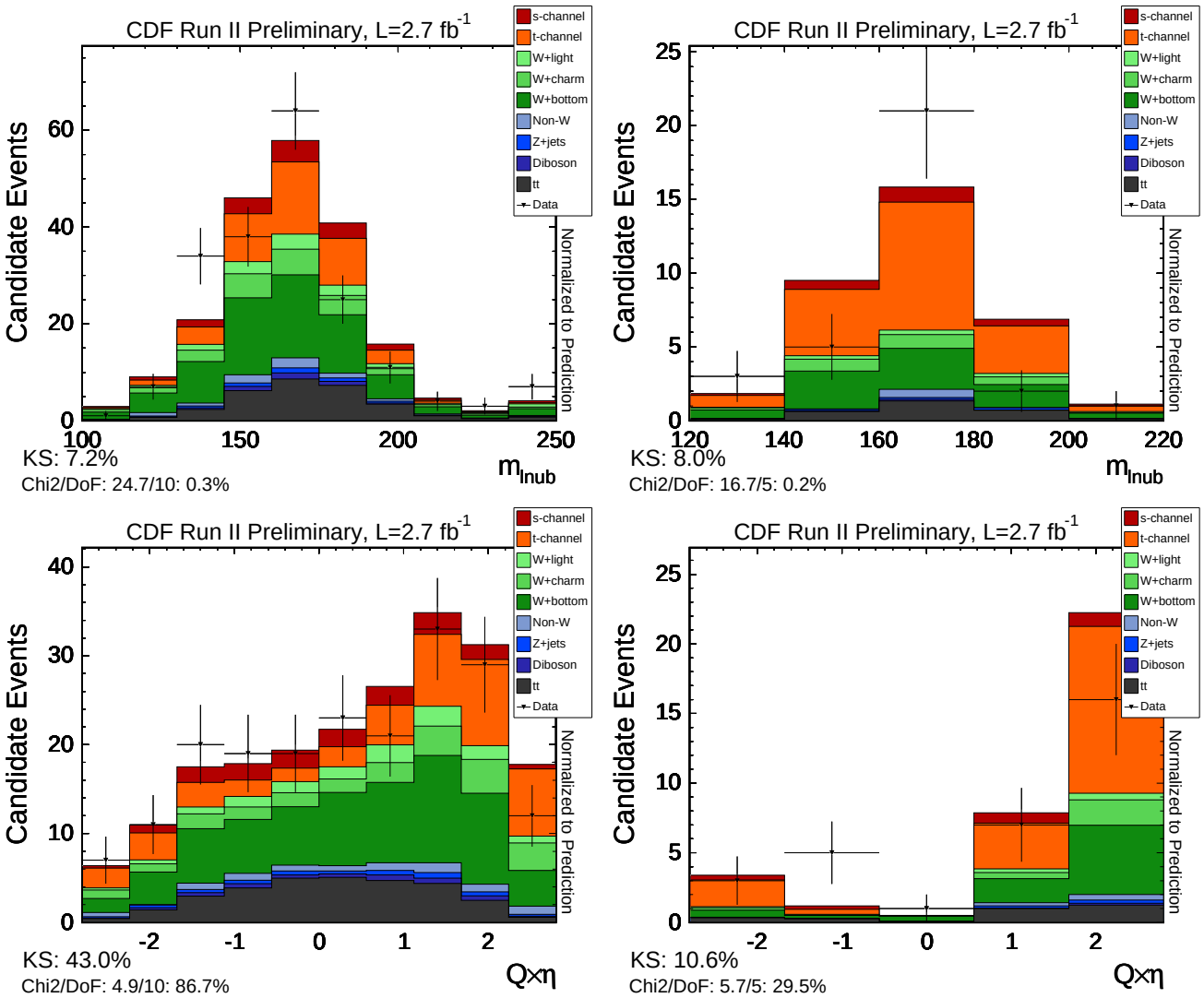

FIG. 7: Data and Monte Carlo comparison of the $Q_{lepton} \cdot \eta_{untagged\ jet}$ and m_{Wb} distributions for increasing cuts on the BDT output. The top row includes events with BDT output values (BDT>0.25) and the bottom row includes events with BDT output values (BDT>0.6).

[1] B.W. Harris *et. al.*, *Phys. Rev.* D **66**, 054024 (2002)
Z. Sullivan *Phys. Rev.* D **70**, 114012 (2004).

[2] G. Mahlon, *Phys. Rev.* D **55**, 7249 (1997), hep-ph/0011349.

[3] T. M. P. Tait and C.-P. Yuan, *Phys. Rev.* D **63**, 014018 (2002).

[4] CDF and DØ Collaborations, FERMILAB-PUB-03/320-E (2003).

[5] F. Caravaglios *et. al.*, M. L. Mangano *et. al.*, JHEP 0307:001 (2003).

[6] L. Demortier, *Bayesian treatments of Systematic Uncertainties*, Proceedings of Advanced Statistical Techniques in Particle Physics, Grey College, Durham, 18 - 22 March 2002

[7] A. Hocker, P. Speckmayer, J. Stelzer, F. Tegenfeldt, H. Voss, K. Voss, A. Christov, S. Henrot-Versille, M. Jachowski, A. Krasznahorkay Jr., Y. Mahalalel, R. Ospanov, X. Prudent, M. Wolter, A. Zemla, *TMVA - Toolkit for Multivariate Data Analysis*, arXiv:physics/0703039v4 (2007)

[8] C. Ciobanu, T. Junk, T. Müller, P. Savard, B. Stelzer, W. Wagner, T. Walter, *Likelihood Function for Single Top Search with 162 pb*$^{-1}$, CDF Note 7106 (2004)

[9] Particle Data Group, W.-M. Yao *et al.*, J. Phys. **G33**, 1 (2006).

[10] L. Read, J.Phys G **28**, 2693 (2002) and T. Junk, Nucl. Instum. Meth. **434**, 435 (1999). See also P. Bock *et al.* (The LEP Collaborations), CERN-EP-98-046 (1998) and CERN-EP-2000-055 (2000).